PSYCHOLINGUISTICS ON A LARGE SCALE: COMBINING TEXT CORPORA, MEGASTUDIES, AND DISTRIBUTIONAL SEMANTICS TO INVESTIGATE HUMAN LANGUAGE PROCESSING

Paweł Mandera

Promotor: Prof. Dr. Marc Brysbaert

Co-promotor: Dr. Emmanuel Keuleers

Proefschrift ingediend tot het behalen van de academische graad van Doctor in de Psychologie

2016

Table of Contents

Acknowledgements

Writing this dissertation would not be possible had it not been for all the help I received along the way. First and foremost, I owe a great deal of gratitude for the generous support I received from my supervisors, Prof. Dr. Marc Brysbaert and Dr. Emmanuel Keuleers. I often felt spoiled by the luxury of having both of you always ready to listen to me.

Marc, I would like to thank you especially for sharing your enthusiasm for research and guiding me so skillfully during these years. Thank you for pushing things gently when they needed to be pushed while always making sure that I am sane, happy and interested in doing the necessary work. You allowed me to directly benefit from your inexhaustible productivity but also thought me a lot about how I can pursue it myself. Thank you for all the freedom, guidance and time you gave me to learn and to develop my ideas.

Emmanuel, since we first met in Kraków almost exactly 5 years ago you have opened many doors for me, for which I will be always grateful. Not only I would never finish writing this dissertation without you but I would have never even started. I would like to thank you for everything you did for me as a supervisor: for all the great discussions, your valuable insights and relentless editing of my writing. Last but not least, I would like to express my gratitude to you for being not only my supervisor but also a great friend. Thank you for all the road-trips, fun at the conferences, and enlightening me about food and wine.

I am grateful to Prof. Dr. Victor Kuperman, Prof. Dr. Antal van den Bosch and Prof. Dr. Rob Hartsuiker who constituted my advisory

committee board. Thank you for taking time to listen to me and always providing thoughtful feedback and encouragement. Your insights and critical comments were extremely appreciated.

All my current and former colleagues at the Department of Experimental Psychology also deserve appreciation. Thank you for making the department such a nice place to work. I really enjoyed its friendly atmosphere. I would like to distinguish Michaël Stevens, with whom I shared an office for all this time. Michaël, thank you for being such a great company. Talking about various things, ranging from statistics over gadgetry to house renovation, always helped me relax during these long days.

I owe special gratitude to Ark Verma and Vatsala Khare for providing me with a roof over my head after I came to Belgium, for telling me so much about being a PhD student in Ghent but also about India and cricket.

I would like to acknowledge everyone who I had a chance to collaborate with during these years. I owe my gratitude to my co-authors, Dr. Zofia Wodniecka and Prof. Dr. Walter Van Heuven. I thank Kevin Tang for teaching me a lot about linguistics and for his relentless enthusiasm and energy that he poured into the projects that we worked on together.

Through their unconditional support my parents always made it feel safe to take up challenges. Thank you for always being there and making me know that I can always count on you. Mateusz, thank you for being so helpful with math and all the enthusiastic discussions we had in the middle of the nights.

Agata, you have been an essential part of all of this. I value your sacrifices. Thank you for keeping me sane, for making me feel loved and simply for being here for and with me on every single day.

Paweł

Ghent, March 2016

Chapter 1. Introduction

When students embark on a dissertation project in psychology, they often have an idealized vision of the enterprise of research. They will get up in the morning, awakened by a creative spark and formulate a research hypothesis that will clearly follow from a well-defined theory. All they have to do next is to design and perform an elegant experiment in the afternoon, analyze the data in the evening, and finish writing a paper by the end of the week.

Of course, this idealized version of the cycle of research does not fully reflect the research practice. The first reason is that the time-frame described above resembles more that of a TV-show in which events occurring in the course of an entire year are compressed into 50 minutes. The second reason is that it lacks a dramatic element, namely the constant worry about *p*-values. The third reason, which is the one I will focus on, is that researchers never function completely independently: A great deal of a their time is spent on a frustrating wait for colleagues or scientists from other fields to deliver the tools and resources that are essential to their research question. In psycholinguistics in particular the researcher depends on many language-related resources, from generally used resources such as dictionaries, grammars, over text corpora, lists of word frequencies, and ratings of stimuli collected from human subjects, to specialized software, to compute lexical statistics or to create nonword stimuli in many languages.

Researchers who work on languages such as English, Dutch, or French are relatively lucky, because the essential materials and tools are relatively well-developed for these languages. In many other

languages, however, a complete lack of reliable resources makes psycholinguistic research almost impossible.

In our idealized example, the researcher does not encounter these practical limitations and has complete freedom in formulating the hypothesis and then designing the experiment. In reality, many researchers will find that some critical resources are missing and be forced to give up and modify their research questions or to provide the materials for themselves. Luckily, the second option is becoming more and more feasible as the ubiquity of storage of information in digital form has caused a rapid increase in the total amount of data one can collect and use for research.

First, it is far easier now to collect text corpora that can be used to create resources such as word frequencies for different languages, different registers, etc. Second, traces of behavior are constantly being produced by anyone using a digital device and, because many of such devices are now connected via Internet, a researcher can even design on-line experiments (e.g., Crump, McDonnell, Gureckis, & Gilbert, 2013) that will easily reach a huge number of potential participants without bringing them to the laboratory.

It has been proposed that developments like these are *transformative* for the science of psychology, leading to a new field which has tentatively been called psychoinformatics, which combines novel data collection techniques and methods from computer science and machine learning with psychology (Yarkoni, 2012; Markowetz, Błaszkiewicz, Montag, Switala, & Schlaepfer, 2014)[1]. Importantly, the

[1] Other scientific fields have seen a similar evolution. For example, in the field of genomics, the increase in the amount of biological data has led to the emergence of bioinformatics, which is now a firmly established field, credited for many of the current advances in biology

mere fact that more data are available and that they can be processed quickly and efficiently, is *not* the only transformation. Instead, these resources and methods also allow researchers to find new ways to *do* psychology.

First, in contrast to data collected in traditional psychological experiments which are designed to answer one specific question, large datasets are conductive to being re-used and re-analyzed. This also encourages sharing the datasets with other researchers, who can often provide a new perspective and deliver new insights based on the same data.[2]

Second, large datasets promote application of analytical techniques that go beyond traditional null-hypothesis testing, which has been strongly criticized for being poorly understood and mindlessly applied (Gigerenzer, 2004). Large datasets encourage the expansion of the psychologist's statistical toolbox. For instance, the mere fact that the large number of data points often provide enormous statistical power and allows to detect even trivial effects naturally shifts researcher's focus to comparisons of effect sizes.

Third, large datasets are suitable for performing exploratory analyses which are considered as an essential part of the scientific process (Jewett, 2005), especially in a field such as psychology, where we are often unable to even formulate the problem in a way that can be solved (O'Donohue, & Buchanan, 2001).

Fourth, easier compilation of psycholinguistic resources offers an opportunity to create resources for *underresourced* languages and

[2] Publicly releasing the data or the resource often has the additional positive side-effect of increasing the number of citations of the associated paper, because the researchers refer to it every time they use the resource or the dataset.

to compile tailor-made resources for studying specific populations. For example, a researcher investigating child-language can easily compile word frequencies based on materials targeted towards children if those materials are available in digital form.

Finally, the increased availability of digital materials and the potential to reach large populations of participants using web-based experiments can remove the practical limitation that often leads to excessive focus on easily accessible groups of participants such as undergraduate students of the university where the researcher works. This may be especially important in psycholinguistics where we need to make sure that the results generalize not only to all demographic groups but also to all languages (see Myers, submitted).

To come full circle: In the age of psychoinformatics, a day in a researcher's life could consist of data-mining on rich, publicly available datasets of behavioral data to develop ideas about the direction in which a theory could evolve. The researcher could then design an experiment based on specialized materials that exactly match their needs and collect data from thousands of participants in a remote location. Alternatively, they could test the hypothesis on another already existing dataset. In the course of the afternoon, they would make the data available to other researchers, who could use it in a computational model or for purposes completely unrelated to the original research. Because collecting data would become so easy that every study would have enough power to actually detect non-trivial effects, this day in the life would unfortunately lack the dramatic element of the constant concern about *p*-values.[3]

[3] To be clear, this idealized version is probably as remote from the current reality as the first one.

What this idealized version of a day in a researcher's life has in common with the first one, is that events spanning the course of a year are compressed into a completely unrealistic timespan. It does however illustrate that psychological research does not consist of doing purely theoretical research in a vacuum. Instead, it encompasses the entire enterprise that makes research possible.

In this dissertation, I deal simultaneously with the development of new resources that are valuable to the field of psycholinguistics and beyond, the improvement of the methodology for developing these resources, and with theoretical questions in the field of lexical processing that can be addressed with these resources.

The common ground of the chapters in this dissertation is that they all consist of the development or exploitation of a new resource, the methodological challenges associated with this development, and reflection on theoretical questions that can be addressed using these resources. In describing the chapters that form the core of this dissertation, I will therefore focus on three different elements: First, what does the research bring to the research community in terms of resources?; Second, how has the research improved methodology for developing new resources?; Third, what are the theoretical questions that were addressed using these resources?

In the first empirical chapter of this dissertation (chapter 2), we present a new set of word frequency norms for British English based on a stream of subtitles broadcasted on BBC channels over a period of two years. The dataset also includes information about frequencies of parts-of-speech and lemmas associated with different words as well as frequencies derived from materials targeted towards children. The quality of the new resource is evaluated by comparing it to the word

frequency norms derived from the British National Corpus. Methodologically, an interesting aspect of this dataset is that the subtitles are encoded as text and contain reliable metadata about the associated programs. This allowed us to compile a corpus which does not suffer from problems with optical character recognition and in which duplicates were easy to detect. Being broadcasted on British television and accompanied by British English captions, the contained materials contrast with the widely-used SUBTLEX-US list word frequencies (Brysbaert & New, 2009), in which American English dominates. Using two lexical decision megastudy databases – one conducted in the USA (Balota et al., 2007) and one conducted in the UK (Keuleers, Lacey, Rastle, & Brysbaert, 2012), it was possible to address the long-standing methodological and theoretical concern to what extent differences between British and American English in the source material for frequency norms have an effect on predicting behavioral data from participants in the UK and in the US .

Chapter 3 moves the focus away from English and presents the first database of word frequencies based on movie subtitles for Polish. Similarly to the resources presented in chapter 2, the database also includes information about the frequencies of parts-of-speech and about lemmas associated with each word form. This information is especially valuable for researchers working in Polish because it is a highly inflected language. The absence of existing behavioral data that could be used to evaluate the word frequencies inspired a methodological question which is important for developing psycholinguistic research in any languages, for which limited resources are available: How can we evaluate frequency norms in the most efficient way. First, I investigate whether the validation of

frequency norms can be made easier by using a carefully selected set of validation stimuli instead of the traditional method which uses large amounts of behavioral data, and therefore requires an existing megastudy. Second, I collect the validation data in a small web-based experiment, which has the additional advantage that participants do not need to be locally available. A theoretically interesting aspect of chapter 3 is that the new resource allowed us to look at the relative importance of the inflected word form frequency and the lemma frequency for predicting performance in the lexical decision task in a highly inflecting language such as Polish.

In chapter 4, I take the approach of web-based experiments a step further by analyzing data from two massive on-line vocabulary tests conducted in Dutch and English with almost 1.5 million participants in total. These experiments were designed primarily to collect a large amount of data regarding word knowledge in a wide population of individuals. The collected datasets have already proved to be a useful source of knowledge about human language processing (Keuleers, Stevens, Mandera, & Brysbaert, 2015; Brysbaert, Stevens, Mandera, & Keuleers, 2016). I analyze the response times collected in these experiments and show that megastudies can be extended from the existing approach, where responses to many stimuli are collected on relatively few participants in laboratory settings, to a new approach in which responses to a large number of stimuli are collected from a very large set of demographically diverse participants using browser-based presentation on a wide range of devices including smartphones and tablets. This is important for two reasons. First, it allows us to evaluate to what extent effects found in frequently studied groups of participants, typically undergraduate students, generalize to other

groups. Second, because the geographical proximity of the researcher and the participants is irrelevant, it removes a boundary in doing psycholinguistic research in languages that are currently understudied. In this chapter, I establish that the chronometric data collected in the aforementioned studies are useful for psycholinguistic research by using a standard psychometric approach to reliability measurement as well as by looking at the qualitative pattern of correlations with existing datasets. I also show that differences in empirical effects between groups of participants in these studies are informative for psycholinguistic theory. Specifically, I show that the changes in the word frequency effect associated with age, proficiency and education can be explained by combining the simple learning principles described with the power function (Newell & Rosenbloom, 1981) with the properties of word frequency distributions.

The first three empirical chapters of the thesis combined megastudy data with corpus-based measures of word frequency, a critical variable for visual word recognition. However, word frequency is just one of the possible measures that can be derived from text corpora. The last two empirical chapters focus on how information derived from text corpora using distributional semantics methods can be useful in psycholinguistics. The most prominent models of this kind, such as HAL (Lund & Burgess, 1996), LSA (Landauer & Dumais, 1997) or Beagle (Jones & Mewhort, 2007), have been used in psycholinguistics for some time. Recent developments have made the application of distributional semantics to psycholinguistics particularly interesting: New methods of deriving semantic information have become available (Mikolov, Chen, Corrado, & Dean, 2013), text corpora have become larger and more specialized, and

large datasets of behavioral data that tap into semantic knowledge have been published (e.g., Hutchison et al., 2013). However, using distributional semantics models often requires specialized knowledge of programming, data processing, and access to substantial computational resources.

In chapter 5, I create distributional semantic spaces for Dutch and English, and present a novel visual interface that allows researchers to explore semantic spaces resulting from the analysis of word co-occurrence data. The interface can also be used with spaces created by other researchers, independently of the underlying model used to generate the spaces. The methodological and theoretical questions in this chapter are how and why traditional approaches to distributional semantics differ from the newer class of models (Mikolov et al., 2013). I address these questions by discussing the theoretical relationships between the different types of the distributional semantics models, by evaluating and discussing their performance in predicting human behavior on a broad set of tasks and by investigating the effect that different text corpora have on their performance. Based on these results, I provide a set of distributional semantic spaces for English and Dutch that should be of particular value in psycholinguistics (i.e., they perform very well on the aforementioned tasks) and that can be used with the interface I developed.

Finally, in chapter 6, I use existing databases of human ratings to move beyond the simple evaluation logic that I employed in chapter 5. This chapter finds its roots in a proposal by Bestgen & Vincze (2012) that distributional semantics models could be used in combination with extrapolation methods to estimate human ratings

based on a small seed of ratings. For example, if one knows that the word *cake* and *party* have positive valence and that the word *birthday* is semantically related to these two, one can make an informed guess about whether the word *birthday* is positive or negative. Following the enthusiasm associated with these methods I hoped that it would be possible to create large datasets for new variables and for underresourced languages. I analyse this approach by examining whether (1) the estimated values can substitute for original ratings in research practice and (2) whether the extrapolation procedure introduces statistical artifacts to the estimated values, which would make it impossible to use these values as a substitute for the original ratings, for instance as an experimental or control variable in behavioral research. I also investigate for which variables and in which way the semantic component is informative, which is theoretically interesting. For example, when applied to age-of-acquisition ratings, this method can give us an idea if language acquisition follows a thematically organized trajectory and whether semantically similar words are acquired around the same age.

REFERENCES

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459.

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods, 44*(4), 998–1006. http://doi.org/10.3758/s13428-012-0195-z

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441–458. http://doi.org/10.1037/xhp0000159

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE, 8*(3), e57410. http://doi.org/10.1371/journal.pone.0057410

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587–606. http://doi.org/10.1016/j.socec.2004.09.033

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., … Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods, 45*(4), 1099–1114. http://doi.org/10.3758/s13428-012-0304-z

Jewett, D. L. (2005). What's wrong with single hypotheses?: Why it is time for Strong-Inference-PLUS. *Scientist, 19*(21), 10.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1–37. http://doi.org/10.1037/0033-295X.114.1.1

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. T*he Quarterly Journal of Experimental Psychology, 68*(8), 1665–92.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203–208.

Markowetz, A., Błaszkiewicz, K., Montag, C., Switala, C., & Schlaepfer, T. E. (2014). Psycho-informatics: big data shaping modern psychometrics. *Medical Hypotheses, 82*(4), 405–411.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. Retrieved from http://arxiv.org/abs/1301.3781

Myers, J. (submitted). Meta-Megastudies. Retrieved from http://www.ccunix.ccu.edu.tw/~lngproc/Myers_MetaMegastudies_DraftShow_151015.pdf

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

O'Donohue, W., & Buchanan, J. A. (2001). The weaknesses of strong inference. *Behavior and Philosophy*, 1–20.

Yarkoni, T. (2012). Psychoinformatics: New Horizons at the Interface of the Psychological and Computing Sciences. *Current Directions in Psychological Science, 21*(6), 391–397. http://doi.org/10.1177/0963721412457362

Chapter 2. SUBTLEX-UK: A new and improved word frequency database for British English[1]

ABSTRACT

We present word frequencies based on subtitles of British television programmes. We show that the SUBTLEX-UK word frequencies explain more of the variance in the lexical decision times of the British Lexicon Project than the word frequencies based on the British National Corpus and the SUBTLEX-US frequencies. In addition to the word form frequencies, we also present measures of contextual diversity part-of-speech specific word frequencies, word frequencies in children programmes, and word bigram frequencies, giving researchers of British English access to the full range of norms recently made available for other languages. Finally, we introduce a new measure of word frequency, the Zipf scale, which we hope will stop the current misunderstandings of the word frequency effect.

[1] This chapter was published as Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190.

INTRODUCTION

Word frequency arguably is the most important variable in word recognition research (Brysbaert, Buchmeier, et al., 2011). Words that are often encountered are processed faster than words that are rarely encountered. Figure 1 shows the course of the word frequency effect. It includes mean standardized reaction times (z-values) for samples of 1000 words going from an average frequency of 0.06 per million words (a $\log_{10}$ value of −1.2) to an average frequency of nearly 1000 per million words (a $\log_{10}$ value of nearly 3.0). The reaction times come from the English Lexicon Project (ELP; circles; Balota et al., 2007) and the British Lexicon Project (BLP; squares; Keuleers, Lacey, Rastle, & Brysbaert, 2012), which contain lexical decision times to over 40 thousand words of American English (ELP) or over 28 thousand monosyllabic and disyllabic words of British English (BLP). The word frequencies come from the British National Corpus (BNC; Kilgarriff, 2006), a 100-million-word collection of samples of mostly written and some spoken language from a wide range of sources, collected between 1991 and 1994 and designed to represent a wide cross-section of British English at that time. Another database of word frequency norms often used for British English is the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995), based on a corpus of 17.9 million words assembled along the same criteria as those for the BNC.

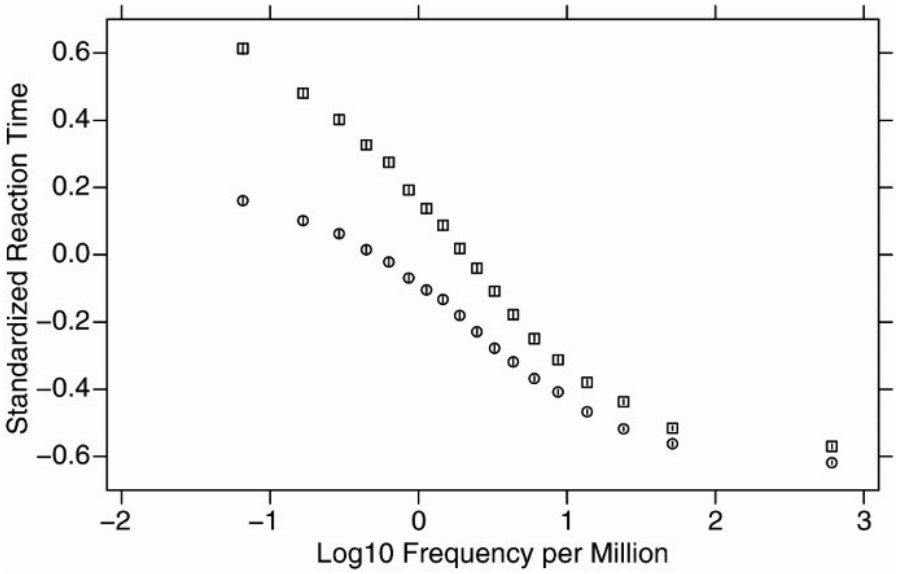

Figure 1 The course of the word frequency effect in mean standarized reaction times from the British Lexicon Project (squares) and the English Lexicon Project (circles). The standard errors are represented by whiskers.

Research in American English and other languages has suggested that word frequencies based on film and television subtitles are better predictors of word processing times than word frequencies based on books and other written sources (Brysbaert, Buchmeier, et al., 2011; Brysbaert, Keuleers, & New, 2011; Brysbaert & New, 2009; Cai & Brysbaert, 2010; Cuetos, Glez-Nosti, Barbon, & Brysbaert, 2011; Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010; Ferrand et al., 2010; Keuleers, Brysbaert, & New, 2010; New, Brysbaert, Veronis, & Pallier, 2007). This is an important finding, because the more variance can be explained by word frequency the fewer other variables are needed to account for word processing times. Brysbaert and Cortese (2011), for example, found that word familiarity did not explain much extra variance in lexical decision

times to monosyllabic English words when the SUBTLEX-US subtitle frequency measure was used (Brysbaert & New, 2009) instead of a commonly used, outdated frequency measure based on a small corpus of written sources (Kučera & Francis, 1967).

Although word frequency estimates based on American subtitles can be used (and have been used) in British word recognition research, some precision is lost, because some words have a different spelling (e.g., labor vs. labour) or a different meaning (e.g., biscuits, pants) in the two languages. The divergences between American and British word usage imply that British researchers should limit their research to the words fully shared among the languages if they use American subtitle frequencies. Otherwise, their findings risk overestimating the impact of nonfrequency variables, such as age of acquisition, word familiarity, word length, or similarity to other words. Suboptimal frequency estimates also increase the risk of stimulus selection errors. This will be the case when words must be selected on the basis of frequency information (e.g., words having different numbers of closely resembling words, so-called orthographic neighbours, with higher frequencies) or when words of different conditions must be matched on frequency (e.g., highly emotional words vs. neutral words).

To address the limitations that researchers working with British English are confronted with, we decided to collect subtitle-based UK word frequency norms. In addition, because we were able to directly capture the subtitles from a variety of television programmes, for the first time we also collected subtitle frequencies from channels specifically aimed at children. Below we describe the collection of the

data, the summary statistics calculated, and the first validation studies we ran.

METHOD

CORPUS COLLECTION

In line with UK regulations, since 2008 the British Broadcasting Corporation (BBC) subtitles all scheduled programmes on its main channels, to help the hearing impaired.[2] These subtitles are not broadcasted through the main channel, but can be superimposed on the programme by those who wish so (e.g., by using Teletext). To have the widest possible range of language input, we collected the words and word pairs of the subtitles from nine channels (BBC1–BBC4, BBC News, BBC Parliament, BBC HD, CBeebies, and CBBC) broadcasted over a period of three years (January 2010–December 2012). Of these channels, BBC1 is the most popular and extensive (aimed at all types of audiences). The other channels have more limited hours. Of further interest is that the CBeebies channel is meant for preschool children (0–6 years) and the CBBC channel for primary school children (6–12 years). This allowed us to compile frequency norms for these groups.

Notwithstanding the provisions relating to "fair dealing" provided under Section 29 of the Copyright Designs & Patents Act 1988 (Government United Kingdom, 1988), the full textual content of the relevant subtitles was not stored or reproduced for the purpose of this research. A count of individual words and consecutive words was

[2] On the basis of anecdotal evidence we can add that these subtitles are also appreciated by viewers with English as second language.

undertaken, obtainable from public transmissions. The method employed does not detract from or otherwise undermine the value of this evaluative work.

TEXT CLEANING

The broadcasts were cleaned semiautomatically for doubles (programme repeats) and subtitle-related information not broadcasted to the viewers. Also the parts of the subtitles not related to the conversation were eliminated (e.g., the words "silence" or "thunder" to describe the ongoing scene; these are usually presented in upper case, or in a different font or colour in the subtitle). After the cleaning we obtained a total of 201.7 million words, coming from 45,099 different broadcasts. This is larger than the other existing subtitle corpora (Brysbaert & New, 2009; Cai & Brysbaert, 2010; Cuetos et al., 2011; Dimitropoulou et al., 2010; Keuleers et al., 2010)[3] and allowed us to calculate more precise parts-of-speech dependent frequencies and word bigrams.

WORD FREQUENCY MEASURES

WORD FREQUENCY COUNTS

A first decision to be made was what to do with hyphenated words. In British English, words are often hyphenated when they function as adjectives. So, a potion that saves lives can be described as "a life-saving potion". This phrase could be counted as consisting of three word types (a, life-saving, potion) or four word types (a, life,

[3] Brysbaert and New (2009) reported that the word type frequencies themselves show little difference once the corpus contains 30 million words, a finding that was replicated in the present analyses.

saving, potion). The problem was particularly relevant for the BBC subtitles, because nearly one out of four word types contained a hyphen in the first analysis of the data. The vast majority of these hyphenated entries were of low frequency (fewer than 100 observations on a total of 200 million words). Because there are no a priori considerations about how to handle this finding (also because there is quite some individual variability in the use of hyphens; Kuperman & Bertram, 2013), we decided to use a pragmatic criterion and looked at which word frequencies correlated most with the 28 thousand lexical decision times of the BLP (Keuleers et al., 2012). As this clearly favoured the dehyphenated word frequencies (a difference in variance explained of 5%), we decided to dehyphenate the data before counting the words.[4]

The dehyphenated subtitles resulted in a total of 332,987 different word types for a total of 201,712,237 tokens. Of these, 31,368 types were in the CBeebies subtitles with a total of 5,860,275 tokens, and 70,755 types were in the CBBC subtitles with a total of 13,644,165 tokens. Because the vast majority of words observed in a single broadcast were typos and other nonword-like structures (like "aaaarrrrgh" or "zzzzzzzzzzzz"), we decided to take out all entries observed in a single broadcast only. This reduced the number of types to 159,235 with a total token count of 201,335,638 for the complete corpus, 5,848,083 for the CBeebies subcorpus (27,236 types), and 13,612,278 for the CBBC subcorpus (58,691 types).

[4] Dehyphenation also occurs in automatic text parsers, such as CLAWS and the Stanford parser (to be described later). Because the Stanford parser dehyphenates more words than CLAWS, the outcome of this parser outperformed that of CLAWS on the raw corpus, but no longer on the dehyphenated corpus.

A STANDARDIZED FREQUENCY MEASURE: THE ZIPF SCALE

Although the frequency counts are the most versatile measure (as will become clear later, when we calculate all types of derived measures), they have one big disadvantage. The interpretation of the frequency measure depends on the size of the corpus. Therefore, authors have looked for a standardized frequency measure, an index with the same interpretation across all corpora collected.

Thus far, the most popular standardized frequency measure has been frequency per million words (fpmw). It is the frequency measure that we made available in our previous work on subtitle frequencies as well. However, we increasingly noticed that this measure leads to an incorrect understanding of the word frequency effect.

Because their corpus contained only 1 million words, the lowest value in the word frequencies made available by Kučera and Francis (1967) was 1 fpmw. This contributed to the assumption that 1 fpmw is the lowest possible frequency. Obviously, this is no longer the case for larger corpora. As it happens, about 80% of the word types in SUBTLEX-UK have a frequency of less than 1 fpmw (i.e., fewer than 200 occurrences in all broadcasts). Second, as shown in Figure 1, nearly half of the word frequency effect is situated below 1 fpmw, and there is very little difference above 10 fpmw. The frequency effect of lexical decision times between 0.1 fpmw and 1 fpmw is equal to or larger than the effect between 1 fpmw and 10 fpmw. A logarithmic transformation of frequency measures, as is routinely performed, alleviates this problem. However, the logarithms of fpmw become negative for frequencies lower than 1 (as again shown in Figure 1), which uninformed users tend to avoid. Because of these properties, fpmw as a standardized measure puts users on the wrong foot.

To make the word frequency effect easier to understand, one needs a scale with the following properties:

1. It should be a logarithmic scale (e.g., like the decibel scale of sound loudness).
2. It should have relatively few points, without negative values (e.g., like a typical Likert rating scale, from 1 to 7).
3. The middle of the scale should separate the low-frequency words from the high-frequency words.
4. The scale should have a straightforward unit.

Once we know what the scale should look like, it is not so difficult to come up with a good transformation. In particular, when we take the $\log_{10}$ of the frequency per billion words (rather than fpmw), the scale fulfils the first three requirements. To meet the last requirement, we propose to call the new scale the *Zipf scale*, after the American linguist George Kingsley Zipf (1902–1950) who first thoroughly analysed the regularities of word frequency distribution and formulated a law (Zipf, 1949), which was later named after him. The unit then becomes the Zipf.

The Zipf scale is a logarithmic scale, like the decibel scale of sound intensity, and roughly goes from 1 (very-low-frequency words) to 6 (very-high-frequency content words) or 7 (a few function words, pronouns, and verb forms like “have”). The calculation of Zipf values is easy as it equals $\log_{10}$ (frequency per billion words) or $\log_{10}$ (frequency per million words) + 3. So, a Zipf value of 1 corresponds to words with frequencies of 1 per 100 million words, a Zipf value of 2 corresponds to words with frequencies of 1 per 10 million words, a Zipf value of 3 corresponds to words with frequencies of 1 per million words, and so on.

Table 1 summarizes the information. It also helps to clear one more misunderstandings about word frequencies among psycholinguists, namely that words with frequencies below 1 fpmw are too uncommon to be known. There are hundreds of derived and inflected word forms and even lemmas with frequencies of lower than 0.1 fpmw that are perfectly known, as can be seen in Table 1. Content words rarely have a Zipf value higher than 6, so that for most practical research purposes, the Zipf scale will be a scale from 1 to 6 with the tipping point from low frequency to high frequency between 3 and 4.

Table 1. The Zipf scale of word frequency

Zipf value	*fpmw*	*Examples*
1	0.01	antifungal, bioengineering, farsighted, harelip, proofread
2	0.1	airstream, doorkeeper, neckwear, outsized, sunshade
3	1	beanstalk, cornerstone, dumpling, insatiable, perpetrator
4	10	dirt, fantasy, muffin, offensive, transition, widespread
5	100	basically, bedroom, drive, issues, period, spot, worse
6	1000	day, great, other, should, something, work, years
7	10000	and, for, have, I, on, the, this, that, you

Note. The Zipf scale is a word frequency scale going from 1 to 7. Words with Zipf values of 3 or lower are low-frequency words; words with Zipf values of 4 and higher are high-frequency words. Examples are based on the SUBTLEX-UK word frequencies. fpmw = frequency per million words.

One more addition that is of interest for the Zipf scale is the possibility to include words with frequency counts of 0 (i.e., words not observed in the corpus). Although these words are less common in large corpora, they are by no means absent. Such words pose a problem for the Zipf scale as a result of the logarithmic transformation (given that the logarithm of 0 is minus infinity). In a recent review,

Brysbaert and Diependaele (2013) concluded that the best way to deal with 0 word frequencies is the Laplace transformation. Rather than working with the raw frequency counts, one works with the frequency counts + 1. This means that all frequency values are (slightly) elevated. The proper application of the algorithm also implies that the theoretical size of the corpus is a little larger than the actual size, because one is leaving room for *N* unobserved word types with frequency 1. *N* is the number of word types in the frequency list. So, for the full corpus the Laplace transformation assumes that there are 159,235 unobserved word types extra in the language, all with a frequency of 1.

In practice, the following equation is needed to calculate the Zipf values on the basis of the frequency counts of the total corpus:

$$Zipf = \log_{10}\left(\frac{\text{frequency\_count} + 1}{201.336 + 0.159}\right) + 3.0$$

The values in the denominator are the size of the corpus in millions plus the number of word types in millions. Specifically, the Zipf value of an unobserved word type will be:

$$Zipf = \log_{10}\left(\frac{0 + 1}{201.336 + 0.159}\right) + 3.0 = 0.696$$

The Zipf value of a word type observed once in the complete corpus will be 0.997; that of a word observed 10 times will be 1.737, and so on.

To calculate the Zipf values for the CBeebies corpus, we have to use the following equation:

$$Zipf = \log_{10}\left(\frac{\text{frequency\_count}_{CBeebies}+1}{5.848+0.027}\right)+3.0$$

For the CBBC subcorpus the equation is

$$Zipf = \log_{10}\left(\frac{\text{frequency\_count}_{CBBC}+1}{13.612+0.059}\right)+3.0$$

Specifically, this means that words with a 0 frequency in the CBeebies corpus get a Zipf value of 2.231; those with a 0 frequency in the CBBC corpus get a Zipf value of 1.864. The higher values for unobserved word types are due to the smaller sizes of the corpora and also mean that one should be sensible in their use. There is no point in blindly using these values for all missing words in the lists, as one assumes that the missing words are known to preschoolers (CBeebies) or primary school children (CBBC). As we see below, this may be one reason why the childhood frequencies are not correlating very well with the lexical decision times of the British Lexicon Project when calculated across all words.

To give readers a better feeling for the Zipf scale, Table 2 tabulates the summary statistics of the Zipf values used in two classic word frequency studies in British English (Monsell et al., 1989; Morrison & Ellis, 1995). Two interesting observations can be made. First, the standard deviations of the Zipf values are similar for the high- and the low-frequency words (as they should be), whereas for fpmw the standard deviations are considerably larger in the conditions with high-frequency words than in the conditions with low-frequency words. Second, we see that in both studies the low-frequency words had Zipf values above 3, because the researchers derived their frequency estimates from the Kučera and Francis list and considered 1

fpmw as the lower end of the frequency range. With the availability of more refined word frequency measures, we hope that in the future more use will be made of words with Zipf values below 3. As Figure 1 indicates, this is a sensible thing to do, as in this range the word frequency effect is at its strongest. Furthermore, about 80% of the word types in SUBTLEX-UK have Zipf values below 3 (i.e., below 1 fpmw). So, there is much more choice at the low end of the distribution than at the high end. In our current estimate, low-frequency words ideally have a mean Zipf value at (or below) 2.5, and high-frequency words have a mean Zipf value of 4.5.

Table 2. Frequencies used in two classical studies of the word frequency effect, expressed as frequency per million words and as Zipf values

Study	*Condition*	*Fpmw*	*Zipf*
Monsell et al. (1989)	Low frequency words (N = 48)	2.12 (2.22)	3.15 (.39)
(Experiments 1 and 2)	Medium frequency words (N = 48)	15.40 (10.81)	4.09 (.29)
	High frequency words (N = 48)	84.65 (62.66)	4.78 (.40)
Morrison & Ellis (1995)	Low frequency words (N = 24)	6.52 (4.61)	3.66 (.44)
	High frequency words (N = 24)	166.03 (168.4)	5.07 (.37)
	Early acquired words (N = 24)	33.49 (34.8)	4.34 (.44)
	Late acquired words (N = 24)	9.91 (16.5)	3.63 (.55)

Note. Means, with standard deviations in parentheses. Frequencies based on SUBTLEX-UK. fpmw = frequency per million words.

CONTEXTUAL DIVERSITY

Adelman, Brown, and Quesada (2006; see also Adelman & Brown, 2008; Perea, Soares, & Comesaña, 2013; Yap, Tan, Pexman, & Hargreaves, 2011) argued that not so much the frequency of occurrence of a word matters, but the number of contexts in which the word appears. Words only encountered in a small number of contexts (say, a word with a frequency of 100 occurring in one or two

television episodes) will be more difficult to process than equally frequent words encountered in a variety of contexts (e.g., a word with a frequency count of 100 used in 80 different broadcasts). A good proxy for contextual diversity (CD) is the number of television programmes/films (or the percentage of programmes/films) in which the word appears. Brysbaert and New (2009) indeed observed that log(CD) explained up to 4% of variance more in lexical decision times than log(frequency). Part of the advantage was methodological, however. Two factors were involved. First, the effect of log(CD) on reaction times (RTs) is more linear than the effect of log(frequency), which becomes flat for high-frequency words, as can be seen in Figure 1. When nonlinear regression analysis was used, the difference between CD and frequency became smaller than 2%. Another part of the difference was due to the fact that some words occurred with very high frequency in a few films because they were the names of main characters (e.g., archer, bay, brown). The CD statistic is less influenced by these instances than the frequency statistic.

Still, the CD measure seems to have added value. Therefore, we provide this information for the different corpora we used (full corpus, CBeebies, CBBC). The values are available both as the total number of television programmes in which the word occurred and as the percentage of television programmes in which the word was encountered. As indicated above, the total number of broadcasts in the complete corpus was 45,099. The number of broadcasts in CBeebies was 4847; in CBBC it was 4848.[5]

[5] The reason why these numbers are very similar is that both channels have a similar rotation of programmes with repeats after a rather short period of time.

PART-OF-SPEECH DEPENDENT FREQUENCIES

For many purposes it is good to know what roles words play in sentences and the relative frequencies of these roles (Brysbaert, New, & Keuleers, 2012). This enables researchers interested in nouns, for instance, to limit their stimulus materials to words that are always (or mostly) used as nouns. It also allows researchers to know whether an inflected word is used more often as an adjective (e.g., appalling) or as a verb (e.g., played). This is important information to decide which words to include in rating studies (e.g., Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012).

Part-of-Speech (PoS) frequencies can only be obtained after the corpus has been parsed (i.e., the sentences broken down into their constituent parts) and tagged (i.e., the words given their correct part of speech in the sentence). For a long time this was virtually impossible given the amount of work involved. However, the development of automatic PoS taggers has made it possible to get a reasonably good (though not perfect) outcome in reasonable time and at an affordable price. For a long time, the CLAWS tagger developed at the University of Lancaster was the golden standard (Garside & Smith, 1997; Lancaster University Centre for Computer Corpus Research on Language, n.d.). It was used for the BNC corpus, and we also used it for our SUBTLEX-US corpus (Brysbaert et al., 2012). However, in recent years the Stanford tagger (initial version: Toutanova, Klein, Manning, & Singer, 2003; The Stanford Natural Language Processing Group, n.d.) has become a worthy competitor. As it happens, the outcome of the first analyses with the Stanford tagger correlated more with the BLP word processing times than the outcome of the CLAWS tagger did. As indicated in Footnote 3, this was due to the fact that the

Stanford tagger is more consistent in dehyphenating words than CLAWS. When the subtitles were cleared of hyphens before running the taggers, both gave comparable output.

Another advantage of the Stanford software[6] is that it gives the most likely lemma associated with an inflected form. The lemmatization is based on an algorithm developed by Minnen, Carroll, and Pearce (2001). It works on two main principles. First, it looks up whether a word form is present in the dictionary. If so, then the associated lemma can be read out. If a word is lacking, the most likely lemma is allocated on the basis of rules and pattern comparisons (e.g., the most likely lemma of the stimulus "martialisations", identified as a noun, is "martialisation"; and the most likely lemma of the stimulus "Martialis", identified as a name, is "Martialis"). As discussed at greater length in Brysbaert et al. (2012), the outcome of these algorithms is not 100% correct[7] and, hence, should always be checked by the user, certainly for low-frequency words. However, they are a big step forward (with accuracy estimates of 97% and higher) and, therefore, are provided in our database. More precisely, we give information about the most frequent PoS associated with each word type, the frequency of this PoS, and the lemma associated with it, next to all the parts of speech associated with the word type and their

[6] A disadvantage of the Stanford tagger is that in its default mode it Americanizes the spellings of the words. So, one must be careful to change this when one is working with British spellings.

[7] A notorious example is "horsefly", which both CLAWS and Stanford parse as an adverb (arguably because the word is not in the programme's lexicon, so that too much reliance is put on the end letters –ly). Ironically, Stanford does correctly classify "horseflies" as a noun associated with the lemma "horsefly" (presumably because the end letters, –lies, are more likely to be associated with plural nouns than with other parts of speech).

respective frequencies. Because of the lemmatization and because the output was as good as that of CLAWS, the data presented in the SUBTLEX-UK database are based on the Stanford parser and tagger. Table 3 gives an example of the output. All frequencies are given as raw frequency counts based on the entire corpus, because this value is the most informative to calculate derived statistics from (e.g., the percentage use as the dominant PoS).

BIGRAM FREQUENCIES

Because extra information can be obtained from word combinations (Arnon & Snider, 2010; Baayen, Milin, Filipovic Durdevic, Hendrix, & Marelli, 2011; Siyanova-Chanturia, Conklin, & van Heuven, 2011), we also collected word bigram frequencies in the entire corpus (i.e., the frequency with which word pairs were observed). This resulted in over 1.5 million lines of consecutive word pairs observed in the corpus. For each pair we give information about the number of times it was observed, the symbols written between the words (space, punctuation mark, hyphen, …) and their respective frequencies. This makes it possible for everyone to calculate interesting additional metrics. For instance, it allowed us to add the 787 hyphenated words with a frequency count of more than 100 (fpwm = 0.5) to the database.[8]

[8] These frequencies were not subtracted from the frequencies of the individual words, under the assumption that the component words of a hyphenated word get coactivated upon seeing the hyphenated word.

Table 3. Example of the PoS analysis

RowNr	*Spelling*	*DomPoS*	*DomPoSLemma*	*DomPoSFreq*	*DomPoSLemmaTotalFreq*	*AllPoS*	*AllPoSFreq*
50277	finalisation	noun	finalisation	5	5	.noun.	.5.
50278	finalise	verb	finalise	164	466	.verb.noun.	.164.6.
50279	finalised	verb	finalise	206	466	.verb.adjective.	.206.5.
50280	finalises	verb	finalise	10	466	.verb.	.10.
50281	finalising	verb	finalise	86	466	.verb.noun.	.86.3.
50282	finalist	noun	finalist	703	2201	.noun.adjective.name.verb.	.703.77.12.2.
50283	finalists	noun	finalist	1498	2201	.noun.name.	.1498.18.
50284	finality	noun	finality	28	29	.noun.	.28.
50285	finally	adverb	finally	27804	27804	.adverb.name.	.27804.2.
50286	finals	noun	final	4450	4450	.noun.name.	.4450.52.
50287	finaly	adverb	finaly	4	4	.adverb.	.4.
50288	finance	noun	finance	3364	3364	.noun.name.verb.	.3364.1225.628.
50289	financed	verb	finance	335	1102	.verb.	.335.
50290	financer	noun	financer	3	4	.noun.	.3.
50291	finances	noun	finances	2806	2806	.noun.verb.name.	.2806.11.1.
50292	financesed	verb	financese	4	4	.verb.	.4.
50293	financess	noun	financess	2	2	.noun.	.2.
50294	financial	adjective	financial	15048	15048	.adjective.name.	.15048.1302.
50295	financialisation	noun	financialisation	3	3	.noun.	.3.
50296	financially	adverb	financially	1557	1557	.adverb.	.1557.
50297	financials	noun	financial	43	43	.noun.	.43.
50298	financier	noun	financier	72	150	.noun.name.	.72.1.
50299	financiers	noun	financier	78	150	.noun.	.78.

Note. PoS = part of speech. For each word type (in the column “Spelling”), the most frequent PoS (DomPoS), the associated lemma (DomPoSLemma), the number of times this PoS is observed in all SUBTLEX-UK subtitles (DomPosFreq), the total frequency of the lemma in the subtitles (DomPoSLemmaTotalFreq), all parts of speech associated with the word type (AllPoS), and the frequencies of these parts of speech in all subtitles (AllPoSFreq) were determined. From this table, we see that according to the Stanford tagger, the word type “finalise” is used mostly (164 times) as a verb (associated with the lemma “finalise”), but also occasionally (6 times) as a noun. The total frequency of the verb lemma “finalise” (which also includes the frequencies of the word types “finalises”, “finalised”, and “finalising”) is 466.

It also allowed us to warn researchers when a compound word is more likely to be written as two separate words than as a single word (for instance, the word “makeup” is observed 308 times in the subtitles (Zipf = 3.18), but the spellings “make-up” and “make up” have a combined frequency of 8998, making “makeup” a bad choice for a low-frequency word).

CORRELATIONS WITH LEXICAL DECISION MEASURES

Given the ease with which word frequencies can be collected nowadays, it is important to check whether a new frequency measure adds something extra to the existing ones. On the basis of previous research, we can expect this to be the case given the superiority of subtitle-based frequency estimates, but still it is good to test this explicitly, also to make sure no calculation errors have been made. The most interesting dataset is the BLP (Keuleers et al., 2012), which provides lexical decision reaction times and accuracy measures of British students for over 28 thousand monosyllabic and disyllabic words. The main competitors to the SUBTLEX-UK word frequencies are the BNC frequencies, the CELEX frequencies, and the SUBTLEX-US frequencies. Words not observed in a corpus were assigned a frequency of 0, and log frequencies were the Zipf values (with Laplace transformation). The Laplace transformation was also used for the CD measure.

Table 4 shows the results for the accuracy data. As expected, the SUBTLEX-UK frequencies outperform the other measures, more so for the CD measure than for the Zipf measure. Because of the large number of observations, the differences are all highly significant. For

instance, the t-value of the Hotelling–Williams test (Steiger, 1980)[9] of the difference in correlation with SUBTLEX-UK (Zipf) and BNC (Zipf) equals 16.8 ($df = 28{,}282$, $p < .001$). In terms of percentage variance explained, the difference is nearly 3%, which is high given that many variables explain less than 1% of variance, once the effects of word frequency, word length, and similarity to other words are partialled out (Brysbaert, Buchmeier, et al., 2011; Brysbaert & Cortese, 2011; Kuperman et al., 2012).

Interestingly, the correlations with the childhood frequencies are much lower, in particular the correlation with the CBeebies frequencies (preschool children). Two reasons for this are the smaller sizes of the corpora (including the many missing words not known to children but given rather high Zipf estimates) and the fact that the overall SUBTLEX-UK frequencies include the subtitles from CBeebies and CBBC television programmes (almost 10% of the total SUBTLEX-UK).

Table 5 shows the correlations for the reaction times (RTs) to the words. Because RTs are only interesting when the words are known, we set percentage accuracy to >66% ($N = 20{,}557$). Very much the same picture appears, with superior performance for the SUBTLEX-UK measures (CD slightly more so than Zipf).

[9] An easy introduction to the test and an Excel file to calculate the exact values are available on the website (http://crr.ugent.be/ archives/546)

Table 4. Correlations between the various frequency measures and the BLP accuracy data

	SUBTLEX-UK	*SUBTLEX-UK_CD*	*SUBTLEX-US*	*BNC*	*Celex*	*CBeebies*	*CBBC*
Accuracy	.600	.628	.557	.564	.553	.390	.535
SUBTLEX-UK		.992	.881	.898	.858	.724	.887
SUBTLEX-UK_CD			.877	.904	.866	.702	.876
SUBTLEX-US				.830	.830	.705	.851
BNC					.927	.633	.789
Celex						.642	.778
CBeebies							.821
Percentage of variance accounted for							
SUBTLEX-UK (Zipf)	40.40%						
SUBTLEX-UK (log(CD+1))	47.10%						
SUBTLEX-US (Zipf)	35.70%						
BNC (Zipf)	35.90%						
Celex (Zipf)	34.60%						

Note. The upper part shows the correlations. The lower part shows the percentages of variance accounted for by nonlinear regression analyses (lm-procedure in R, restricted cubic splines with 4 knots). BLP = British Lexicon Project; BNC = British National Corpus; CD = contextual diversity. N = 28,285.

Table 5. Correlations between the various frequency measures and the BLP RT data

	SUBTLEX-UK	*SUBTLEX-UK_CD*	*SUBTLEX-US*	*BNC*	*Celex*	*CBeebies*	*CBBC*
RT	-.664	-.674	-.645	-.638	-.624	-.535	-.642
SUBTLEX-UK		.991	.885	.900	.862	.727	.893
SUBTLEX-UK_CD			.878	.906	.869	.701	.880
SUBTLEX-US				.822	.828	.698	.847
BNC					.937	.611	.771
Celex						.626	.762
CBeebies							.817
Percentage of variance accounted for							
SUBTLEX-UK (Zipf)	46.1%						
SUBTLEX-UK (log(CD+1))	47.1%						
SUBTLEX-US (Zipf)	43.3%						
BNC (Zipf)	42.2%						
Celex (Zipf)	40.7%						

Note. The upper part shows the correlations. The lower part shows the percentages of variance accounted for by nonlinear regression analyses (lm-procedure in R, restricted cubic splines with 4 knots). BLP = British Lexicon Project; BNC = British National Corpus; CD = contextual diversity; RT = reaction time. N = 20,557.

Table 6. Percentages of variance accounted for by the various frequency measure in the ELP data

	Accuracy_LDT (N = 40,468)	*RT_LDT (N = 33,997)*	*RT_nam (N = 33,997)*
SUBTLEX-US (Zipf)	20.5%	36.7%	26.0%
SUBTLEX-US (CD)	22.3%	37.2%	26.1%
SUBTLEX-UK (Zipf)	19.0%	34.8%	24.2%
SUBTLEX-UK (CD)	20.5%	34.8%	24.2%

Note. ELP = English Lexicon Project; CD = contextual diversity; RT = reaction time; LDT = lexical decision task.

Table 7. Correlations of the SUBTLEX-UK frequencies with the CPWD word frequencies

Frequency measure	*SUBTLEX-UK (Zipf)*	*CBeebies (Zipf)*	*CBBC (Zipf)*
CPWD	.664	.756	.690
SUBTLEX-UK (Zipf)		.734	.925
Cbeebies (Zipf)			.803

Note. All values log transformed after Laplace transformation; N = 9125 word types shared between both lists. CPWD = Children's Printed Word Database.

To make sure that the higher correlations between SUBTLEX-UK and the BLP measures than between SUBTLEX-US and BLP were due to language congruency and not to the better quality of SUBTLEX-UK overall, we ran similar analyses of the ELP data, which were collected on American students. As can be seen in Table 6, the difference between SUBTLEX-UK and SUBTLEX-US indeed has to do with differences in word use between the two languages rather

than with the inherent qualities of the frequency lists. Whereas the SUBTLEX-UK frequencies are better for the British BLP data (see Tables 4 and 5), the SUBTLEX-US data are better for the American ELP data (Table 6).

CORRELATIONS WITH THE CHILDREN'S PRINTED WORD DATABASE (CPWD)

The best existing British database of word frequencies for children is the Children's Printed Word Database (CPWD; available at http://www.essex.ac.uk/psychology/cpwd/; checked on May 21, 2013). It includes the frequencies with which 12,193 different word types are observed in 1011 books (995,927 tokens) for 5–9-year-old children in the UK (Masterson, Stuart, Dixon, & Lovejoy, 2010). We could download data for 9659 word types from the database, 9125 of which were also in the SUBTLEX-UK list (the ones not in the list were mainly genitive forms, hyphenated forms, and numbers). Table 7 gives the correlations between log CPWD frequencies and various SUBTLEX-UK frequencies for the 9125 shared word types. As can be seen, the correlations are reasonably high, in particular with the CBeebies word frequencies. The Hotelling–Williams test indicated significant differences between the CBeebies frequencies and the other frequencies (e.g., difference between CBeebies and CBBC, $t(9122) = 15.6$, $p < .001$). This confirms that the SUBTLEX-UK children frequencies are an interesting addition to the CPWD frequencies and can be used to study frequency trajectories from childhood to adulthood[10] (Lété & Bonin, 2013).

[10] SUBTLEX-UK frequencies not including childhood frequencies can easily be obtained by subtracting the CBeebies and CBBC frequency counts from the total frequency counts.

Discussion

In this paper, we presented a new database of word frequencies for British English, based on television subtitles. On the basis of our previous research, we expected that these frequencies would better predict word processing performance than word frequencies based on written sources (in particular, the British National Corpus). This indeed turned out to be the case, when we tried to predict the lexical decision times and accuracies of the British Lexicon Project (Tables 4 and 5). The British subtitle frequencies were also better for predicting the BLP data than were the American subtitle frequencies, but they were inferior for accounting for the ELP data, in line with the observation that word usage is not completely the same in British and American English. The extra variance accounted for amounted to 3–5%, which is considerable given that many variables explain less than 1% of the variance once the effects of word frequency, length, and similarity to other words are partialled out (Brysbaert, Buchmeier, et al., 2011; Brysbaert & Cortese, 2011; Kuperman et al., 2012).

While analysing the findings, we were once again struck by how misleading the standardized word frequency measure fpmw (frequency per million words) is to understand the word frequency effect. Therefore, we proposed an alternative, the Zipf scale, which is better suited to the use of word frequencies in psychological research. This scale goes from slightly less than 1 to slightly more than 7 and can easily be interpreted as follows: Values of 3 and less are low-frequency words; values of 4 or more are high-frequency words. Words not in SUBTLEX-UK get a Zipf value of 0.696 when the frequencies are based on the complete corpus, 1.864 when the CBBC frequencies are used, and 2.231 when the CBeebies frequencies are

used. The differences in minimal values are caused by the differences in corpus size and agree with the fact that missing words of interest in CBeebies or CBBC are likely to be more familiar than words not found in the entire corpus.

In addition to the word frequencies, the new database offers other information, which will allow British researchers to do cutting-edge investigations. These are:

- Part-of-speech-related frequencies, which make it possible for researchers to better control their stimulus materials.
- A measure of contextual diversity (CD), which is particularly interesting for predicting which words will be known and which not (compare Tables 4 and 5).
- Word frequencies in materials aimed at very young (preschool) and young (primary school) children.
- Information about word bigrams.

AVAILABILITY

The SUBTLEX-UK data are available in three easy-to-use files. The first one (SUBTLEX-UK_all) is a 332,988 × 15 matrix containing information of all word types (including numbers) encountered in the dehyphenated subtitles. The 15 columns give information about:

- The spelling of the word type (Spelling).
- The number of times the word has been counted in all subtitles (Freq).
- The number of times the word started with a capital (CapitFreq).

- The percentage of broadcasts containing the word type in all subtitles (CD).
- The number of broadcasts containing the word in all subtitles (CDCount).
- The most frequent part of speech of the word (DomPoS).
- The number of times this dominant Pos was observed (DomPosFreq).
- The lemma associated with the dominant Pos (DomPosLemma).
- The number of times this lemma was observed in all subtitles (DomPosLemmaFreq).
- The summed frequencies of all the times this lemma was observed irrespective of the PoS (DomPosLemmaTotalFreq).
- All parts of speech taken by the word type (AllPos).
- The respective frequencies of these PoS (AllPosFreq).
- The associated lemma information (AllLemmaPos, AllLemmaPosFreq, AllLemma PosTotalFreq).

The second file (SUBTLEX-UK) contains more information about the 160,022 word types (159,235 single words and 787 hyphenated words) that are observed in more than one broadcast and which only contain letter information (i.e., no digits or nonalphanumerical symbols). This file is the file most psycholinguistic researchers will want to use. It has 27 columns, containing:

- The word type.
- The frequency counts in all subtitles, the CBeebies subtitles, the CBBC subtitles, and the British National corpus.

- The Zipf values associated with the various frequencies.
- The CD counts and percentages in the three SUBTLEX corpora.
- The dominant PoS, its associated lemma, and their frequencies.
- All the PoS and frequencies of the word.
- The frequency of the word starting with a capital.
- Whether the lower-case spelling of the word type was accepted by a UK word spell checker (UK), a US word spell checker (US), both spell checkers (UK US), or none (X)[11]. This is an interesting column when words must be selected, and one wants to avoid the inclusion of names or other uninteresting entries.
- Whether the entry contains a hyphen (cf. the 787 added entries with hyphens).
- Whether the entry has another homophonic entry. This is interesting for finding homophones, but also to make sure selected low-frequency words do not have a higher frequency spelling alternative.
- Whether or not the word type has been encountered as a bigram in the subtitles.
- The frequency of the bigram (summed across all types of intervening symbols, in particular, blank spaces, punctuation marks, and hyphens).

Finally, the third file (SUBTLEX-UK_bigrams) contains information about word pairs. Because this file has nearly 2 million lines of information, it cannot be made available as an Excel file

[11] The speller was the MS Office 2007 spellchecker, augmented with a list of lemmas one of the authors (M.B.) is compiling.

(although we have such a file with all entries observed 12 times or more). Each line contains information about Word 1 and Word 2, the frequency of the combination, the CD count of the combination, and which symbols were found between the two words with which frequencies. This is important information when researchers want to include transition probabilities in their investigations, or when expressions (e.g., object names, particle verbs) consist of two words.

SUPPLEMENTAL MATERIAL

Supplemental files are available via the 'Supplemental' tab on the article's online page (http://dx.doi.org/10.1080/13506285.YEAR.850521). They can also be downloaded from our websites (http://crr.ugent.be/, or http://www.psychology.nottingham.ac.uk/subtlex-uk/), where we in addition intend to make them available as online consultable internet databases.

REFERENCES

Adelman, J. S., & Brown, G. D. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review, 115(1)*, 214-227.

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word naming and lexical decision times. *Psychological Science, 17*, 814–823.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language, 62(1)*, 67-82.

Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P., & Marelli, M. (2011), An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*, 438-482.

Baayen, R. H., & Piepenbrock, R. Gulikers. L.(1995). *The CELEX lexical database [CD-ROM]*. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*, 445-459.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A.M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology, 58*, 412-424.

Brysbaert, M. & Cortese, M.J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology, 64*, 545-559.

Brysbaert, M., & Diependaele, K. (2013). Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods, 45*, 422-430.

Brysbaert, M., Keuleers, E., & New, B. (2011). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology, 2, 27*.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new

and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977-990.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding Part-of-Speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods, 44*, 991-997.

Cai, Q. & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLOS ONE, 5,* e10729.

Children's Printed Word Database. Database of printed word frequencies as read by children aged between 5 & 9. Retrieved May 21, 2013, from http://www.essex.ac.uk/psychology/cpwd/

Cuetos, F., Glez-Nosti, M., Barbon, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicologica, 32*, 133-143.

Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-based word frequencies as the best estimate of reading behavior: The case of Greek. *Frontiers in psychology, 1*(218), 1-12.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Meot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*, 488-496.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 102–121). London: Longman.

Government United Kingdom (1988). Copyright, Designs and Patents Act 1988. Retrieved from http://www.legislation.gov.uk/ukpga/1988/48/contents

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new frequency measure for Dutch words based on film subtitles. *Behavior Research Methods, 42*, 643-650.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44*, 287-304.

Kuperman, V., & Bertram, R. (2013). Moving spaces: Spelling alternation in English noun-noun compounds. *Language and Cognitive Processes*, (ahead-of-print), 1-28.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30 thousand English words. *Behavior Research Methods, 44*, 978-990.

Lancaster University Centre for Computer Corpus Research on Language (n.d.). CLAWS part-of- speech tagger for English. Retrieved May 17, 2013, from http://ucrel.lancs.ac.uk/claws/

Lété, B., & Bonin, P. (2013). Does frequency trajectory influence word identification? A cross-task comparison. *The Quarterly Journal of Experimental Psychology, 66(5)*, 973-1000.

Masterson, J., Stuart, M., Dixon, M., & Lovejoy, S. (2010). Children's printed word database: Continuities and changes over time in children's early reading vocabulary. *British Journal of Psychology, 101(2)*, 221-242.

Minnen, G., Carroll, J., & Pearce, D. (2001). Applied morphological processing of English. *Natural Language Engineering, 7(3)*, 207-223.

Monsell, S., Doyle, M.C., & Haggard, P.N. (1989). Effects of frequency on visual word recognition tasks - Where are they? *Journal of Experimental Psychology: General, 118*, 43-71.

Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of experimental psychology. Learning, memory, and cognition, 21(1)*, 116-133.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics, 28*, 661-677.

Perea, M., Soares, A. P., & Comesaña, M. (2013). Contextual diversity is a main determinant of word identification times in young readers. *Journal of Experimental Child Psychology, 116,* 37-44.

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011). Seeing a Phrase" Time and Again" Matters: The Role of Phrasal Frequency in the Processing of Multiword Sequences. *Journal of Experimental Psychology-Learning Memory and Cognition, 37(3)*, 776-784.

Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.

The Stanford Natural Language Processing Group (n. d.). Stanford Log-linear Part-Of-Speech Tagger. Retrieved May 17, 2013, from http://nlp.stanford.edu/downloads/tagger.shtml

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 173-180). Association for Computational Linguistics.

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review, 18(4)*, 742-750.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, Massachusetts: Addison-Wesley

Chapter 3. SUBTLEX-PL: Subtitle-based word frequency estimates for Polish[1]

ABSTRACT

We present SUBTLEX-PL, Polish word frequencies based on movie subtitles. In two lexical decision experiments, we compare the new measures with frequency estimates derived from another Polish text corpus that includes predominantly written materials. We show that the frequencies derived from the two corpora perform best in predicting human performance in a lexical decision task if used in a complementary way. Our results suggest that the two corpora may have unequal potential for explaining human performance for words in different frequency ranges and that corpora based on written materials severely overestimate frequencies for formal words. We discuss some of the implications of these findings for future studies comparing different frequency estimates. In addition to frequencies for word forms, SUBTLEX-PL includes measures of contextual diversity, part-of-speech-specific word frequencies, frequencies of associated lemmas, and word bigrams, providing researchers with necessary tools for conducting psycholinguistic research in Polish. The database is freely available for research purposes and may be downloaded from the authors' university Web site at http://crr.ugent.be/subtlex-pl.

[1] This chapter was published as Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). SUBTLEX-PL: Subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, *47*(2), 471-483.

INTRODUCTION

Word frequency estimates derived from film and television subtitles have proved to be particularly good at predicting human performance in behavioral tasks. Since lexical decision latencies are particularly sensitive to word frequency (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004), correlating human performance in this task with various word frequency estimates became a standard method of validating their usefulness. Word frequencies derived from subtitle corpora were shown to outperform estimates based on written texts for French (New, Brysbaert, Veronis, & Pallier, 2007), English (Brysbaert & New, 2009), Dutch (Keuleers, Brysbaert, & New, 2010), Chinese (Cai & Brysbaert, 2010), Spanish (Cuetos Vega, González Nosti, Barbón Gutiérrez, & Brysbaert, 2011), German (Brysbaert et al., 2011), and Greek (Dimitropoulou, Duñabeitia, Avilés, Corral, & Carreiras, 2010).

Following these developments, we present SUBTLEX-PL, a new set of psycholinguistic resources for Polish, which includes frequency estimates for word forms, associated parts of speech, and lemmas. To our knowledge, this it is the first subtitle word frequency validation study for a Slavic language. In terms of number of speakers, Polish is the largest language in the West Slavic group and the second largest of all Slavic languages after Russian (Lewis, Simons, & Fennig, 2013). It is a highly inflected language and, as compared with most Germanic languages, has a much richer inflection of nouns, adjectives, verbs, pronouns, and numerals. Polish is written in the Latin alphabet, with several additional letters formed with diacritics. In contrast to English, Polish has a transparent orthography:

In most cases, letters or their combinations correspond to phonemes of spoken Polish in a consistent way.

Even though the collection of text corpora of considerable size is easier than ever before, the standard way of validating the quality of the word frequencies based on these corpora has typically involved collection of data for thousands of words in strictly controlled laboratory settings (Balota et al., 2007; Keuleers, Diependaele, & Brysbaert, 2010; Keuleers, Lacey, Rastle, & Brysbaert, 2011). In order to compare frequency estimates derived from two corpora, it may be more efficient to use words for which the two corpora give diverging estimates, rather than a random set of words. This idea is based on the observation that the words for which the frequency estimates between two corpora differ most are also the sources of potential difference in performance of these frequency norms when predicting behavioral data. This approach can increase the statistical power of the experiment; if only randomly sampled words are included in the study, due to very high correlation between different frequency estimates, it is more difficult to detect differences in performance of these estimates without including a very large number of words in the experiment. Dimitropoulou et al. (2010) approached this problem by using a factorial design in which the critical conditions included words with a high frequency in one corpus and a low frequency in the other. In the present study, we will use an approach based on continuous sampling over the full range of word frequencies.

Although using words for which the two corpora give the most diverging estimates may help to detect differences between their performance in predicting behavioral data, there is a possibility that this approach may bias the experiment in favor of one of the

frequency estimates. For instance, words in the formal register tend to have a much higher frequency in written corpora than in spoken corpora. Stimulus selection based solely on a criterion of maximum divergence would lead to a large selection of words from the formal register, while the formal register may represent just a small part of the corpus. To account for this possibility, in Experiment 1, we included an additional set of words that were randomly sampled from all word types observed in the compared corpora. In Experiment 2, we included only randomly sampled words.

CURRENT AVAILABILITY OF FREQUENCY NORMS FOR POLISH

For a long time, the only available word frequency norms for Polish were based on a corpus compiled between 1963 and 1967 (containing about 500,000 words) and published by Kurcz, Lewicki, Sambor, Szafran, and Woroniczak (1990). More recently, several other Polish text corpora have been compiled, and resources such as concordances and collocations have been made available to researchers. This is the case for the IPI PAN Corpus of about 250 million words (Przepiórkowski & Instytut Podstaw Informatyki, 2004), the Korpus Języka Polskiego Wydawnictwa Naukowego PWN (n.d.), containing about 100 million words, and the PELCRA Corpus of Polish (~100 million words; http://korpus.ia.uni.lodz.pl/). To our knowledge, none of them provides an easily accessible list of word frequencies.

The largest of the Polish corpora contains over 1.5 billion words (National Corpus of Polish [NCP]; Przepiórkowski, 2012). It is based mainly on press and magazines (~830 million tokens), material downloaded from the Internet (~600 million tokens), and books (~100

million tokens). It also contains a small sample of spoken, conversational Polish (~2 million tokens). In addition to the full corpus, a significant effort has been invested in creating a subcorpus that is representative of the language exposure of a typical native speaker of Polish. This balanced subcorpus (BS–NCP) contains about 250 million words. Spoken materials (conversational and recorded from media) constitute about 10 % of the subcorpus. The remaining 90 % is based on written texts (mainly from newspapers and books).

Since the word frequencies derived from the NCP balanced subcorpus seem to be the most appropriate existing word frequencies for psycholinguistic research in Polish, we decided to compare them with the new SUBTLEX-PL frequencies.

SUBTLEX-PL

CORPUS COMPILATION, CLEANING, AND PROCESSING

We processed about 105,000 documents containing film and television subtitles flagged as Polish by the contributors of http://opensubtitles.org. All subtitle-specific text formatting was removed before further processing.

To detect documents containing large portions of text in languages other than Polish, we first calculated preliminary word frequencies on the basis of all documents and then removed from the corpus all files in which the 30 most frequent types did not cover at least 10 % of a total count of tokens in the file. Using this method, 5,365 files were removed from the corpus.

Because many documents are available in multiple versions, it was necessary to remove duplicates from the corpus. To do so, we first

performed a topic analysis using Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003), assigning each file to one of 600 clusters. If any pair of files within a cluster had an overlap of at least 10 % unique word-trigrams, the file with the highest number of hapax legomena (words occurring only once) was removed from the corpus, since more words occurring once would indicate more misspellings.

After removing duplicates, 27,767 documents remained, containing about 146 million tokens (individual strings, including punctuation marks, numbers, etc.), out of which 101 million tokens (449,300 types) were accepted as correctly spelled Polish words by the Aspell spell-checker (http://aspell.net/; Polish dictionary available at ftp://ftp.gnu.org/gnu/aspell/dict/pl/) and consisted only of legal Polish, alphabetical characters. All words were converted to lowercase before spell-checking. Because Aspell rejects proper names spelled with lowercase, this number does not include proper names.

FREQUENCY MEASURES

Word frequency

In addition to raw frequency counts, it is useful for researchers to have measures of word frequency that are independent of corpus size. First, we report word frequencies transformed to the Zipf scale[2] (van Heuven, Mandera, Keuleers, & Brysbaert 2014). The Zipf scale was proposed as a more convenient scale on which word frequencies may be measured. In order to reflect the nature of the frequency effect,

---

[2] $z_i = \log_{10}\left(\frac{c_i+1}{\sum_{k=1}^{n} c_k + n} + 9\right)$ (van Heuven, Mandera, Keuleers, & Brysbaert, 2014) where z_i is a Zipf value for word i, c_i is its raw frequency, and n is the size of the vocabulary.

it is a logarithmic scale (like the decibel scale of sound intensity), but, in contrast to the logarithm of frequency per million words, it does not result in negative values for corpora of up to 1 billion words. In order to make interpretation of the frequency values easier, the middle of the scale separates low-frequency from high-frequency words, and, for a majority of words, the measure takes a value between 1 to 7, which resembles a Likert scale. Another compelling property of the Zipf scale is that it allows assigning a value to words that were not observed in a corpus by incorporating Laplace smoothing, as recommended by Brysbaert and Diependaele (2013); without the transformation, such words pose a problem, since the logarithm of 0 is undefined, which makes it impossible to estimate $\log_{10}$ of word frequency per million for these words. In addition to the raw frequency and the Zipf scale frequencies, we also provide the more traditional logarithm of frequency per million words.

Contextual diversity

Adelman, Brown, and Quesada (2006) proposed that the number of contexts in which a word appears may be more important than word frequency itself and that the number of documents in which a word occurs may be a good proxy measure for the number of contexts (contextual diversity [CD]). According to this view, even words with equal frequency would be processed faster if they occur in more contexts. Brysbaert and New (2009) observed that CD accounts for 1%–3% more variance than does word frequency.

Part-of-speech-specific frequencies

For languages with a rich inflectional system, such as Polish, it is crucially important to provide researchers with information above

the level of individual word forms. For each word in SUBTLEX-PL, we also provide the lemma and the dominant part of speech and their frequencies.

Providing the lemma associated with each given word form allows us to group inflected forms of the same word. This may be useful when investigating the specific contributions of surface and lemma frequencies in word processing (Schreuder & Baayen, 1997) or in order to avoid including inflections of the same word when creating a stimulus set for an experiment.

Information about the dominant part of speech allows researchers to choose words of a particular grammatical class (e.g., when a researcher wants to include only nouns in a stimulus list).

To obtain part-of-speech and lemma information for words, we used TaKIPI, a morphosyntactic tagger for Polish (Piasecki, 2007) supplied with the morphological analyzer Morfeusz (Woliński, 2006). The resulting tag set was too detailed for our purposes, so we translated the original tags to a simpler form that includes only information about parts of speech and discards other details.[3] The tagging process assigned each of the word forms consisting of legal Polish alphabetical characters and accepted by the spell-checker to 1 of 78,361 lemmas.

Bigram frequencies

Although in this article we focus on unigram frequencies, we also provide frequency estimates for word bigrams, which are of increasing interest to researchers (Arnon & Snider, 2010; Siyanova-Chanturia, Conklin, & van Heuven, 2011).

[3] For mapping between original and simplified tags, see supplementary materials.

EXPERIMENT 1

METHOD

Stimuli

We selected stimuli from the list of words common to both BS–NCP and SUBTLEX-PL.[4] All stimuli considered for selection contained only alphabetical characters and occurred without an initial capital in most cases. We used the list of 1-grams (available at http://zil.ipipan.waw.pl/NKJPNGrams) to generate the BS–NCP frequency list used in the present study. We processed the raw list by summing frequencies of all forms that were identical after removing punctuation marks attached to some of the forms in the original list.

To make the experiment maximally informative, we chose stimuli for which BS–NCP and SUBTLEX-PL gave highly divergent frequency estimates. We performed a linear regression on the SUBTLEX-PL frequencies, using the BS–NCP frequencies as a predictor. All frequencies were transformed to the Zipf scale. We then ordered the words according to their residual error and chose 155 words from both extremes of the resulting list, ensuring that different forms of the same lemma were not selected more than once. Words at one extreme (with a large positive residual error value) were much more frequent in SUBTLEX-PL than would be expected on the basis of BS–NCP, while words at the other extreme (with a large negative residual error value) occurred much less often in SUBTLEX-PL than would be expected on the basis of BS–NCP. In addition, we randomly

[4] A nonfinal version of SUBTLEX-PL, based on nearly 50 million tokens, was used when choosing stimuli for the experiment.

sampled 155 words from the remaining words, with the probability of each word being selected equal to its probability in the subtitle corpus.

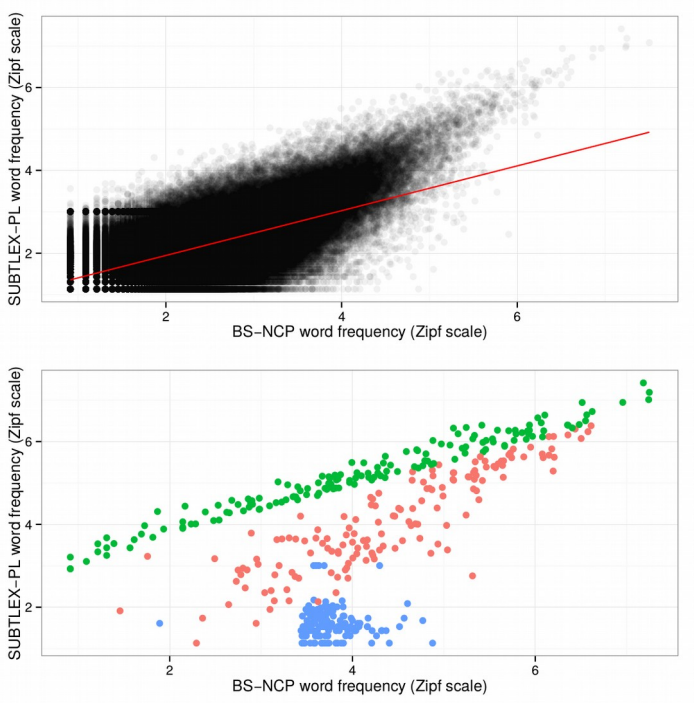

Figure 1. Frequencies of words in the BS–NCP and SUBTLEX-PL corpora for all words (upper panel; the red line shows a regression line predicting SUBTLEX-PL frequencies based on BC–NCP frequencies) and words included in Experiment 1 (bottom panel) showing randomly sampled words (red) and words with higher frequency (green) and lower frequency (blue) in SUBTLEX-PL than in BS–NCP

Figure 1 illustrates the frequency distribution of stimuli according to this procedure. As the top panel of Fig. 1 shows, it is important to note that the regression line on which the residual error values are based is pulled downward by a large number of words with

a low frequency in SUBTLEX-PL. While this seems to indicate that SUBTLEX-PL contains a higher proportion of low-frequency forms, it is an artifact of selecting words from corpora of unequal size.[5]

Words that had a much higher frequency in one corpus than in the other may be categorized into several groups. For example, words related to the Polish administrative and legislative system (e.g., "województwo," *district*; "urzędowym," *administrative*), as well as those occurring mostly in fairly sophisticated contexts (e.g., "pejzażu," *landscape*) are much more frequent in the BS–NCP corpus. On the other hand, words with much higher frequency in SUBTLEX-PL included those used mostly in dialogues (e.g., "skarbie," *honey*), swear words ("pierdol," *fuck*), those related to (American) film themes (e.g., "kowboju," *cowboy*), and function words (e.g., "ale," *but*; "się," *self*).

For each word that was included in the experimental set, a corresponding nonword was generated using Wuggy, a multilingual pseudoword generator (Keuleers & Brysbaert, 2010).

For the full set of words included in the experiment, the standard deviation (*SD*) in word frequency (Zipf scale) was 1.14 (mean = 4.09) for BS–NCP and 1.76 (mean = 3.63) for SUBTLEX-PL. The two variances were significantly different, $F(464, 464) = 0.42$, $p < .001$, and Welsch's *t*-test has shown significant differences in the mean frequency derived from the two corpora, $t(794) = 4.7$, $p < .001$, for this set of stimuli.

[5] As an example, consider a list of 200,000 words and a list of 400,000 words. A typical characteristic of word frequency distributions is that about half of the words in each list will have a frequency of one. In that case, the base probability that any word found in both lists would have a frequency of 1 in the first list would be 1/100,000, while it would be 1/200,000 for the second list.

For the 155 word stimuli that were randomly sampled from the words common to both word frequency lists, *SD* was 1.08 (mean = 4.44) for BS–NCP and 1.19 (mean = 4.11) for SUBTLEX-PL. The difference between variances was not statistically significant, $F(154, 154) = 0.82$, $p = .23$, but the mean frequencies were significantly different according to Welsch's *t*-test, $t(308) = 2.6$, $p = .01$.

Participants

Twenty-six students from the Jagiellonian University in Kraków participated in the experiment (20 female, 6 male; mean age = 23.76, *SD* = 2.06) either on a voluntary basis or in exchange for course credit.

Design

Words and nonwords were randomly assigned to 10 blocks. Nine blocks contained 50 words and 50 nonwords in a random order; 1 block contained the remaining 15 words and nonwords in a random order. Ten different permutations of block orders were generated, and each participant was randomly assigned to one of the permutations.

Due to a coding error, 10 words were not presented to the first 10 participants. Further analysis is therefore based on 455 words, instead of 465 words.

Within each block, stimuli were presented in a random order in white characters on a black background. Presentation of each stimulus was preceded by a blank screen. After 500 ms, a vertical line was displayed above and below the center of the screen. Finally, after another 500 ms, the stimulus was presented between the vertical lines.

A standard QWERTY PC keyboard was used to collect responses. Participants were instructed to press "/" (the rightmost key

on the second row) if they saw a word and "Z" (the leftmost key on the second row) if they saw a nonword. The time-out for giving the response was 2,000 ms. After six training trials, the experimental blocks were presented. The experiment took about 30 min.

RESULTS

Of the trials on which reaction times (RTs) were outside of a range of whiskers of a boxplot adjusted for skewed distributions (calculated separately for words and nonwords for each participant in each block; Hubert & Vandervieren, 2008), 5.2% were removed from the data set.

Accuracy and RTs were the two dependent variables in all analyses. Three stimuli with less than one-third correct answers were excluded from the data set. The analyses are reported first for the full set of words included in the experiment and then separately only for the 155 word stimuli that were randomly sampled from the words common to both word frequency lists.

For the full set of word stimuli, the mean RT was 592.00 (*SD* = 67.34), and the mean accuracy was .94 (*SD* = .08). Words occurring less often in SUBTLEX-PL than in BS–NCP had a mean RT of 652.19 (*SD* = 52.23) and a mean accuracy of .96 (*SD* = .06), while words occurring more often in SUBTLEX-PL than in BS–NCP had a mean RT of 551.02 (*SD* = 48.74) and a mean accuracy of .91 (*SD* = .11). The randomly selected words had a mean RT of 574.00 (*SD* = 54.00) and a mean accuracy of .96 (*SD* = .07).

For nonwords, the mean RT was 666.88 (*SD* = 70.23), and the mean accuracy was .94 (*SD* = .09).

To estimate the reliability of the RT and accuracy measures, we computed split-half correlations for 100 random splits of the data across participants. The resulting correlations were corrected with the Spearman–Brown prediction formula (Brown, 1910; Spearman, 1910), giving an average corrected reliability of .81 (*SD* = .013) for RTs and .72 (*SD* = .021) for accuracy.

Adjusted R^2 was used as a measure of explained variance in all analyses. The percentage of variance in RT and accuracy accounted for by linear regression models using different frequency measures is summarized in Table 1. All frequency measures were transformed to the Zipf scale (van Heuven et al., 2014). Because it was shown that the frequency effect is not completely linear (Balota et al., 2004), we added a term with squared word frequency (Zipf scale) to the linear regression. To control for word length, we also included number of letters in a word in the regression model.

The relationship between word frequencies and RTs is shown in Fig. 2. As is shown in Table 1, when all words were included in the analysis, the BS–NCP word frequencies explained 39.09 % of variance in RTs and 8.90 % of variance in accuracy. For this set of words, SUBTLEX-PL frequencies explained 58.64 % of variance in RTs and 19.07 % in accuracy, which is 19.55 % more for RTs and 10.17 % more for accuracy in comparison with BS–NCP frequencies. To test for statistical difference between models, we applied the Vuong test for nonnested models (Vuong, 1989). The differences in performance of the two models were statistically significant for both RTs ($z = -6.11$. $p < .001$) and accuracy ($z = -2.5$, $p = .012$).

Table 1. Percentages of variance accounted for by the various frequency measures in Experiment 1

Model	*RT (%; all words)*	*Accuracy (%; all words)*	*RT (%; sampled words)*	*Accuracy (%; sampled words)*
$length + WF_{BS-NCP} + WF_{BS-NCP}^2$	39.09	8.90	45.53	20.58
$length + WF_{SUB-PL} + WF_{SUB-PL}^2$	58.64	19.07	53.88	18.43
$length + CD_{SUB-PL} + CD_{SUB-PL}^2$	59.72	20.81	54.35	19.26
$length + WF_{SUB-PL} + WF_{SUB-PL}^2 + DLF$	58.80	20.16	53.59	18.52
$length + CD_{SUB-PL} + CD_{SUB-PL}^2 + DLF$	59.77	21.64	54.10	19.20
$length + WF_{SUM} + WF_{SUM}^2$	50.99	19.14	51.01	22.01
$length + WF_{AVG} + WF_{AVG}^2$	58.36	21.38	55.46	21.77

Note. Columns 2 and 3 show the results for all words in the experiment; columns 4 and 5 show the results for randomly sampled words. WF = word frequency (Zipf scale), DLF = log 10 of dominant lemma frequency, BS–NCP =Balanced Subcorpus–National Corpus of Polish, SUB-PL = Polish Subtitle Corpus, WF SUM = normalized (Zipf scale) sum of word frequencies in SUBTLEX-PL and BS–NCP, WF AVG = averaged Zipf scale frequency in the two corpora

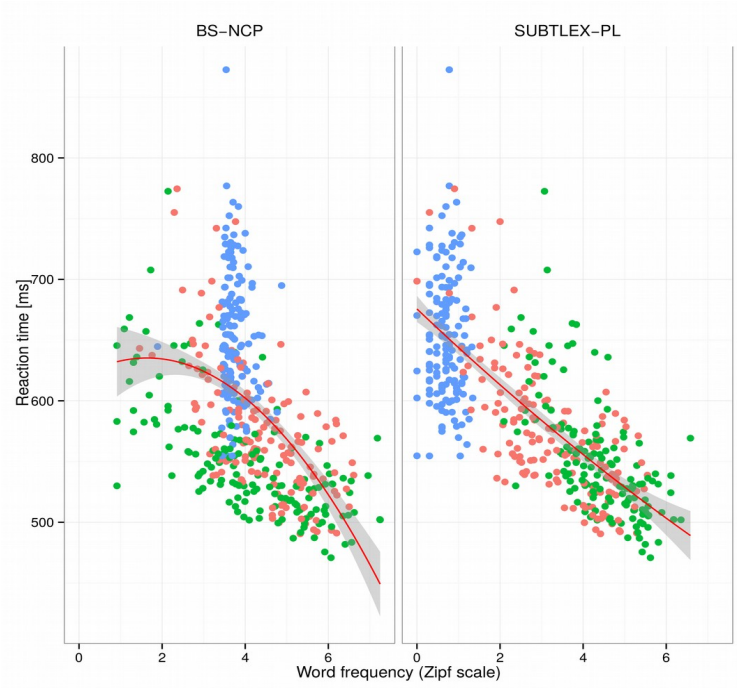

Figure 2. Reaction times in Experiment 1 for words and their frequencies in the BS–NCP (left) and SUBTLEX-PL (right) corpora. Reaction times for words that had much higher frequencies in BS–NCP, as compared with SUBTLEX-PL (blue), are shifted upward from the regression line, while words that have higher frequencies in SUBTLEX-PL than in BS–NCP (green) tend to be responded to faster than would be predicted on the basis of BS–NCP frequencies. Reaction times predicted on the basis of SUBTLEX-PL line up much closer to the regression line. For words that were randomly sampled from the full set of words (red), this difference is less apparent, but it is still reflected in R^2. Red lines represent predictions of a linear model with word frequency and its square term as predictors (with standard error in the shaded area)

When only words that were randomly sampled from the corpus were included in the analysis, the frequencies derived from the BS–NCP corpus explained 45.53 % of the variance in RTs and 20.58 % in accuracy. In this case, the difference between the BS–NCP and

SUBTLEX-PL corpora was smaller, and word frequencies derived from the SUBTLEX-PL corpus explained 8.35 % more variance for RTs but 2.15 % less variance for accuracy. The difference was not significant for RTs ($z = -1.84$, $p = .065$) or accuracy ($z = 0.45$, $p = .65$).

For the full set of words, CD measures calculated on the basis of SUBTLEX-PL accounted for the largest part of the variance for both RTs and accuracy, explaining 59.72 % and 20.81 % of variance, respectively. This improvement of model predictions, relative to the one based on word frequencies, was statistically significant for both RTs ($z = 2.41$, $p = .016$) and accuracy ($z = 2.57$, $p = .010$).When only randomly selected words were included in the analysis, CD explained 54.35 % of variance for RTs and 19.26 % for accuracy. This was not significantly better than the model based on subtitle word frequencies for RTs ($z = 0.86$, $p = .39$) or for accuracy ($z = 1.15$, $p = .25$).

To examine the importance of lemma frequency, we conducted further analyses including dominant lemma frequency as an additional predictor. This predictor turned out to add very little to the total amount of explained variance. The Vuong test has not indicated in any case that the model including this predictor should be preferred over a simpler model.

In addition to analyses based on frequencies derived from SUBTLEX-PL and BS–NCP, we also calculated compound measures of word frequency, taking into account frequencies in the two corpora simultaneously: their summed frequency (transformed to the Zipf scale after summation) and their averaged normalized (Zipf scale) frequency. In the case of the full set of word stimuli, in comparison with BS–NCP frequencies, the summed frequency measure explained

11.89 % more variance in RTs (z= 6.38, p < .001) and 10.24 % more variance in accuracy (z = 2.97, p = .003). In comparison with subtitle frequencies, it explained 7.66 % less variance in RTs (z = −2.93, p = .003) and a similar amount of variance in accuracy (z = 0.016, p = .99). The averaged frequency explained 7.3% more variance in RTs than did the summed frequency (z = 4.40, p < .001) and a comparable amount of variance to subtitle frequencies (z = −0.16, p = .87). For accuracy, its predictions were not significantly better than summed frequencies (z = 0.84, p = .40) or subtitle frequencies (z = 1.03, p = .30) and outperformed only BS–NCP-based frequencies (by 12.50% of explained variance; z = 4.157, p < .001).

For a randomly sampled set of words, the compound measures performed particularly well: The model using estimates based on averaged normalized frequency in the two corpora accounted for 1.1% more variance in RTs than did the next best model (based on SUBTLEX-PL contextual diversity), but the difference between the two models was not statistically significant (z = 0.38, p = .70). In comparison with the model based on BS–NCP word frequencies, both summed frequency (z = 2.86, p = .004) and averaged frequency (z = 3.65, p < .001) performed significantly better in predicting RTs. As compared with the model based on SUBTLEX-PL frequencies, the difference was not statistically significant for either of the compound measures (for summed word frequency, z = −0.073, p = .46; for averaged word frequency z = 0.57, p = .57). The two compound measures were also best at predicting accuracy, but none of the differences in accuracy reached the level of statistical significance (z < 1.96).

DISCUSSION

In Experiment 1, we found a general advantage of SUBTLEX-PL frequencies. The difference was larger when stimuli with extremely divergent frequency estimates were included in the analyzed data set. At first sight, these results suggest that the SUBTLEX-PL word frequencies are more balanced than the BS–NCP word frequencies: RTs for the three different groups of stimuli are in line with the predictions from SUBTLEX-PL. On the other hand, the BS–NCP frequencies seem to severely underestimate RTs for words that have a much lower occurrence in SUBTLEX-PL (shown in blue in Fig. 2). This could indicate that the BS–NCP corpus has inflated frequency estimates for these words, of which most could be characterized as belonging to a very formal register.

However, we should note that the frequency range of the sample of words for which BS–NCP makes the worst predictions is very restricted, making a general conclusion about the global suitability of the BS–NCP frequencies premature. Researchers will not often encounter a situation where an experiment requires exactly this register of words. Moreover, when only randomly sampled words were included in the data set, the difference between performance of the two frequency estimates was smaller, and the advantage of SUBTLEX-PL was no longer statistically significant.

In additional analyses, we have shown that compound frequency estimates, taking into account both corpora simultaneously, can be particularly good predictors of performance in a lexical decision task. This can be due to the fact that considering the two corpora simultaneously involves a significant increase in the overall

size of a sample of a language on which frequency estimates are based. In addition to that, compounding word frequency estimates may help reduce bias for certain registers that may be present in the individual corpora.

In Experiment 2, we propose a comparison of the two word frequency measures in which the entire frequency distribution is examined and undue bias from a particular register is avoided.

EXPERIMENT 2

METHOD

Participants

For the second experiment, 43 female participants and 15 male participants took part in an online experiment. Mean age of the participants was 27.07 (*SD* = 4.08; 1 of the participants did not give information about age).

Stimuli

Three hundred word stimuli were selected using a two-step sampling procedure. First, simple Good-Turing Smoothing (e.g., Gale & Sampson, 1995) was applied to the word frequencies from BS–NCP and SUBTLEX-PL (Brysbaert & Diependaele, 2013). Words that were present in both word frequency lists and had a length of at least three letters were considered for further selection if they were included in the PWN dictionary (http://sjp.pwn.pl). The probability of a word being selected for the experiment was proportional to its simple Good-Turing Smoothed probability, averaged over BS–NCP and SUBTLEX-PL. Once a word had been selected, other words forms of

the same lemma were ignored, avoiding including different inflections of the same word in the stimulus list. Three hundred nonwords were generated using Wuggy (Keuleers & Brysbaert, 2010) on the basis of an independent sample of words from the SUBTLEX-PL and BS–NCP corpora.

Figure 3 shows the relationship between the BS–NCP and SUBTLEX-PL word frequencies for the stimuli in Experiment 2. Standard deviation in word frequency (Zipf scale) was 1.46 (mean = 3.81) for BS–NCP and 1.59 (mean = 3.72) for SUBTLEX-PL. There were no statistically significant differences between frequencies derived from the two corpora in means (Welsh's *t*-test), $t(594) = -0.74$, $p = .46$, or their variances, $F(299, 299) = 1.2$, $p = .14$.

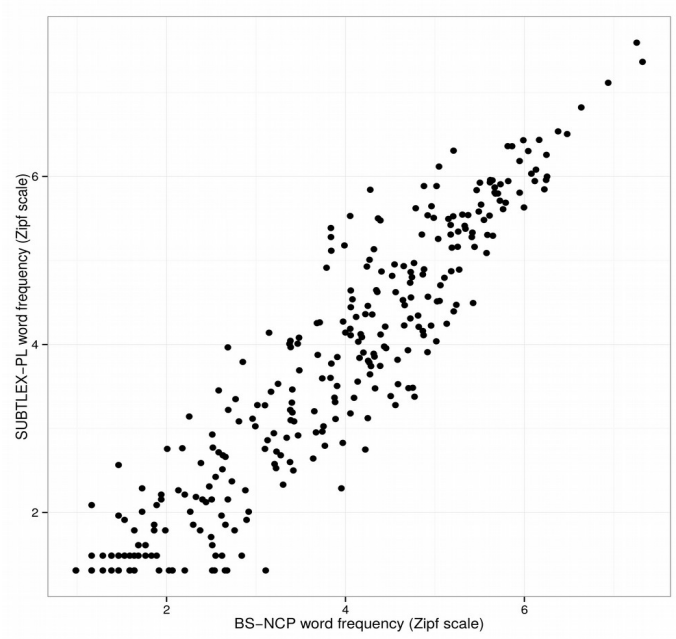

Figure 3. Frequencies in the BS–NCP and SUBTLEX-PL corpora for words included in Experiment 2

Design

The experiment was administered in a Web browser, using custom-designed software, taking into account timing (Crump, McDonnell, & Gureckis, 2013). Participants were instructed to respond by pressing "J" if they thought that the presented stimulus was a word and "F" if they thought that it was not a word. After a short training block with 4 words and 4 nonwords, during which feedback was given after each trial, experimental stimuli were presented in five blocks. For each block, 60 words and 60 nonwords were chosen at random. After each block, feedback was given about performance (mean RT for words and overall accuracy in the preceding block). Participants were allowed to take a short break between blocks. Stimuli were presented in black font on a white background until the participant gave a response, after which the screen would be blank for 500 ms before the next stimulus was displayed. During the experiment, a continuous progress bar was presented in the upper part of the screen.

RESULTS

To exclude outliers from the analyzed data set, a two-step procedure was applied. First, we excluded all trials with RTs longer than 3,000 ms. Next, all observations in which RTs were outside a range of whiskers of a boxplot adjusted for skewed distributions (calculated separately for words and nonwords for each participant in each block; Hubert & Vandervieren, 2008) were removed from the data set. In total, 8 % of trials were removed.

The mean accuracy was .96 for words and .97 for nonwords. Mean RT was 893.97 (*SD* = 188.03) for words and 1,043.79 (*SD* =

174.63) for nonwords. On average, the RTs were substantially longer than in the first experiment, most likely because of the lack of a time-out and the fact that most participants in Experiment 1 were used to taking experiments for course credit.

Reliability of the RT and accuracy measures was computed in the same way as for Experiment 1. The mean corrected reliability was .94 (*SD* = .005) for RTs and .88 (*SD* = .013) for accuracy.

In Experiment 2, as compared with SUBTLEX-PL frequencies, the BS–NCP frequencies accounted for 2.4 % more variance in RTs and for 3 % more variance in accuracy (see also Table 2 and Fig. 4); however, the difference in performance of the two models was not statistically significant for RTs ($z = 1.12$, $p = .26$) or for accuracy ($z = 1.00$, $p = .32$). The compound frequency estimates turned out to give the most accurate predictions of RTs. Although, in comparison with the model based on BS–NCP word frequencies, this difference was not statistically significant for summed frequencies ($z = 1.49$, $p = .14$) or for averaged frequencies ($z = 0.83$, $p = .40$), in comparison with the model based on movie subtitles, both compound measures performed significantly better: The summed frequencies explained 3.4 % more variance ($z = 2.02$, $p = .043$) and averaged frequencies 3.2 % more variance ($z = 2.66$, $p = .008$) in RTs. The model, which included dominant lemma frequencies in addition to subtitle frequencies, significantly outperformed the model without this predictor ($z = 2.11$, $p = .035$).

For accuracy, the measures derived from BS–NCP followed these based on SUBTLEX-PL contextual diversity and dominant lemma frequency in explained percentage of the variance. None of the

differences in accuracy reached the level of statistical significance ($z < 1.96$).

Table 2 Percentages of variance accounted for by the various frequency measures in Experiment 2

Model	*RT (%)*	*Accuracy (%)*
$length + WF_{BS-NCP} + WF_{BS-NCP}^2$	70.48	19.05
$length + WF_{SUB-PL} + WF_{SUB-PL}^2$	68.06	16.02
$length + CD_{SUB-PL} + CD_{SUB-PL}^2$	68.32	17.40
$length + WF_{SUB-PL} + WF_{SUB-PL}^2 + DLF$	70.71	18.96
$length + CD_{SUB-PL} + CD_{SUB-PL}^2 + DLF$	70.72	19.55
$length + WF_{SUM} + WF_{SUM}^2$	71.45	18.37
$length + WF_{AVG} + WF_{AVG}^2$	71.31	18.51

Note. WF = word frequency, BS–NCP = Balanced Subcorpus–National Corpus of Polish, SUB-PL = Polish Subtitle Corpus

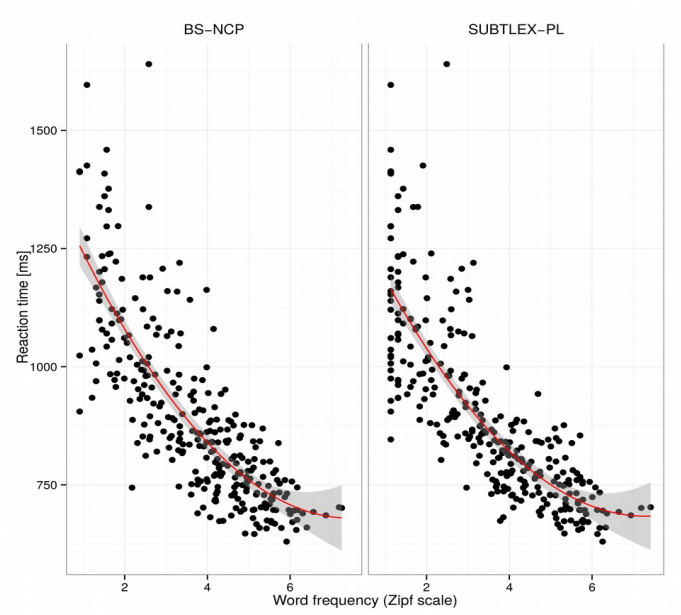

Figure 4. Reaction times for words and their frequencies in the BS–NCP (left) and SUBTLEX-PL (right) corpora. The red lines represent predictions of a linear model with word frequency and its square term

DISCUSSION

In Experiment 2, the compound measures again performed best in predicting behavioral data. Interestingly, for models based on frequency estimates derived from BS–NCP and SUBTLEX-PL, we observed a reversed pattern, relative to Experiment 1: The SUBTLEX-PL frequencies were now worse at predicting RTs, as compared with the compound measures, but this was not the case for BS–NCP frequencies. Even more surprisingly, the randomly sampled words in Experiment 1 showed the reverse pattern. We suspected that this was caused by different means and standard deviations in frequencies between the two experiments. The average frequency was higher in the first experiment (for both corpora) than in the second experiment. Hence, the two corpora may differ in their potential to explain variance in RTs in various frequency ranges. To test this hypothesis, we performed an additional analysis using a linear regression model with number of letters, word frequency in BS–NCP, word frequency in SUBTLEX-PL, and the interaction between the frequencies of both corpora. Table 3 shows the results of this analysis. Because the interaction between the two frequency measures turned out to be highly significant, we decided to conduct an additional analysis. We split the set of words in Experiment 2 at the median point of average word frequency in the two corpora (3.8, Zipf scale). We observed (see Table 4) that the BS–NCP frequencies are better in predicting RTs and accuracy in the lower part of the frequency range, while SUBTLEX-PL frequencies are better in predicting these variables in the higher part of the frequency range. The difference in performance of the models based on frequencies derived from individual corpora was not significant in the upper part of the frequency range ($z = 1.72$, $p = .086$)

or in the lower part of the frequency range ($z = 1.34$, $p = .18$), but the model based on averaged frequencies was best in both frequency ranges. It significantly outperformed BS–NCP-based frequencies in the higher range ($z = 2.34$, $p = .019$) and the model based on subtitle frequencies in the lower range ($z = 2.03$, $p = .042$). For accuracy, the Vuong test did not show preference for any of the models ($z < 1.96$).

In order to verify whether a similar interaction between frequency estimates derived from primarily written-text and subtitle-based corpora can be found in other languages, we conducted an additional analysis using RTs collected in the British Lexicon Project (BLP; Keuleers et al., 2011). We used frequency estimates from the British National Corpus (BNC; Kilgarriff, 2006), which consists mostly of written language and contains about 100 million words, and SUBTLEX-UK (van Heuven et al., 2014). To emulate the setup of the experiment reported in the present article and to better balance the number of words from different frequency ranges, we ran 1,000 simulations in which we randomly chose 300 words from the BLP with weights proportional to the averaged word frequency (Zipf scale) of the BNC and SUBTLEX-UK. For each sample, we fitted a linear model with number of letters, word frequency in the BNC, word frequency in SUBTLEX-UK, and the interaction between the word frequencies of both corpora.

Table 3. Regression model for predicting reaction times using length of a word, frequencies derived from BS–NCP and SUBTLEX-PL, and inter-action term between the two corpora

	Estimate	*SE*	*t-value*	*p*
Intercept	772.67	21.07	36.67	< 2e-16
Length	15.46	2.58	5.99	6.10E-009
$WF_{BS\text{-}NCP}$	-105.87	13.28	-7.97	3.50E-014
WF_{SUB-PL}	-101.47	14.75	-6.88	3.60E-010
$WF_{SUB-PL} * WF_{BS\text{-}NCP}$	17.05	2.61	6.54	2.70E-010

Adjusted R^2 = .71; $F(4, 295) = 186.00$, $p < 2e\text{-}16$

Note. The frequencies were centered before being entered into the linear regression

Table 4. Percentage of variance explained by frequency estimates derived from the two corpora (the data set from Experiment 2 was split at the median)

Frequency	*Model*	*RT (%)*	*Accuracy (%)*
> median	$length + WF_{BS-NCP} + WF_{BS-NCP}^2$	27.49	9.65
> median	$length + WF_{SUB-PL} + WF_{SUB-PL}^2$	33.89	11.72
> median	$length + WF_{SUM} + WF_{SUM}^2$	31.89	11.63
> median	$length + WF_{AVG} + WF_{AVG}^2$	33.79	12.17
<= median	$length + WF_{BS-NCP} + WF_{BS-NCP}^2$	45.70	14.05
<= median	$length + WF_{SUB-PL} + WF_{SUB-PL}^2$	38.38	12.92
<= median	$length + WF_{SUM} + WF_{SUM}^2$	46.20	13.89
<= median	$length + WF_{AVG} + WF_{AVG}^2$	45.45	14.38

We found that the interaction between the two frequency measures was highly significant ($p < .001$) in all 1,000 simulations. At the same time, we did not find an advantage of BNC word frequencies in the lower part of the frequency spectrum when the stimuli in each of the samples was split at the median point (mean median point = 3.21, *SD* = 0.061, Zipf scale). Across all the samples, in the lower part of the range, SUBTLEX-UK frequencies accounted for 9.59 % of the variance (*SD* = 5.00), and BNC frequencies for 6.61 % (*SD* = 4.23) of the variance. In the upper part of the frequency range, SUBTLEX-UK frequencies accounted for 29.73 % (*SD* = 6.9) of the variance, and BNC frequencies for 24.59 % (*SD* = 6.26) of the variance. Interestingly, averaged word frequency accounted for slightly more variance than did SUBTLEX-UK in both lower (mean = 10.53 %, *SD* = 5.00) and upper (mean = 30.25 %, *SD* = 6.58) ranges. The averaged word frequency was also slightly better at predicting RTs for the full set of words (mean = 44.04 %, *SD* = 4.74) than were individual frequency measures (SUBTLEX-UK, mean = 43.23 %, *SD* = 4.77; BNC, mean = 40.34 %, *SD* = 4.66). We compared R^2 values obtained in the simulations using the Welsh *t*-test. Due to the large number of simulations, all reported differences were statistically significant, except for the difference between averaged word frequencies and SUBTLEX-UK frequencies in the upper part of the frequency range.

CONCLUSIONS

We presented new word frequency estimates for Polish based on film and television subtitles and, in two lexical decision experiments, validated their usefulness by comparing them with

estimates derived from BS–NCP, as well as with compound frequency estimates derived from the two text corpora.

We found a large advantage of SUBTLEX-PL over BS–NCP when words for which estimates given by the two corpora differed most were used as stimuli. In contrast, when we sampled words randomly, the advantage became less pronounced (Experiment 1) or tended to favor the BS–NCP-derived frequencies (Experiment 2).

These results suggest that the relationship between frequency estimates derived from different corpora and human performance in behavioral tasks may be complex. In particular, this shows that the stimulus selection procedure may affect the outcome of a validation experiment. For a comparative study to be informative, it is essential to find an unbiased method of stimulus selection. Although it is reasonable to assume that the more words included in a validation study, the more relevant its results, it has to be taken into account that even selecting words from a megastudy for validation (e.g., Keuleers et al., 2010) may introduce bias and make it easier for one of the corpora to provide good frequency estimates than do other corpora. For instance, if only mono- and disyllabic words are included in a study, the mean frequency may be shifted, relative to the mean in the full lexicon, because of a negative correlation between word frequency and word length. In such a case, a corpus that does better in predicting behavioral measures in higher parts of the frequency range would be favored. Using the BLP data, we failed to replicate the advantage of a written text corpus in the lower frequency range, although we found a similar overall interaction between word frequency measures. Also, the small total amount of explained variance in the range below the median point in this analysis may

suggest that mono- and disyllabic words do not represent the lexicon well in that frequency range.

Moreover, it should be considered whether including a full set of words in validation studies is an optimal choice. If a word frequency distribution of a full lexicon were reflected in a stimulus set of a validation study, due to properties of a Zipfian distribution, the vast majority of words would have to be on the low extreme of the possible frequency range, and, because in linear regression all observations contribute equally to the results, R^2 would be determined mostly in the very low part of the frequency distribution. In this case, the results of linear regression would not be very informative for high-frequency words.

Table 5. Regression model for predicting reaction times in Experiment 2 using word length, word frequency (WF_{SUB-PL}), log_{10} of dominant lemma frequency (DLF), and the interaction between form and lemma frequencies

	Estimate	*SE*	*t-value*	*p*
Intercept	787.87	20.69	38.09	< 2E-016
Length	13.52	2.53	5.34	1.90E-007
WF_{SUB-PL}	-137.22	14.26	-9.62	< 2E-016
DLF	-107.38	12.69	-8.46	1.20E-015
$WF_{SUB-PL} * DLF$	21.66	2.91	7.44	1.10E-012

Adjusted R^2 = .719; $F(4, 295) = 193.00$, $p < 2e\text{-}16$

Note. The frequencies were centered before being entered into the linear regression

In addition to these methodological aspects, we would like to point out that it is also possible that some properties of the lexicon may have contributed to the pattern of results obtained in the present

study. It is possible that during word processing, lemma frequency is a source of facilitation that is stronger for low-frequency words than for high-frequency words. As Table 5 shows, in an exploratory analysis, we observed a statistically significant interaction between word frequency and lemma frequency when these two variables and word length were entered into a linear regression as predictors and RTs obtained in Experiment 2 as a dependent variable. It is possible that this extra facilitation for low-frequency words corresponds to slightly higher frequency estimates for low-frequency words in written text corpora than in subtitle corpora. If that were the case, the advantage of the written text corpus, in comparison with the subtitle corpus observed in the low-frequency range, could be incidental, rather than reflecting a real advantage of written-text corpora.

To fully explore these issues, it would be necessary to conduct analyses across different sets of stimuli and for different languages. Lexical decision megastudies (Balota et al., 2007; Keuleers et al., 2010; Keuleers et al., 2011) provide a good opportunity for such analyses.

Nevertheless, even with a validation using a limited set of words, the results of the two experiments suggest that both SUBTLEX-PL and BS–NCP are valuable sources of word frequency estimates. In most cases, we would advise researchers to use the averaged compound measure derived from the two corpora whenever possible. At the same time, we do not have enough evidence to strongly suggest the same practice in other languages. It must also be kept in mind that for certain classes of words, one of the corpora may give strongly biased frequency estimates. We have shown that for BS–

NCP, a subset of low-frequency words used mostly in formal communication may belong to such a category.

AVAILABILITY

SUBTLEX-PL frequencies and compound SUBTLEX-PL/BS–NCP frequencies are available for research purposes and can be downloaded in RData and csv formats from http://crr.ugent.be/subtlex-pl. They can also be accessed online using a Web interface. Frequencies for words with contextual diversity above 2 are also available in the xlsx (Microsoft Excel) format.

The whole word frequency data set for individual words is contained in two files. The first file includes all strings found in the text corpus with rich information about their part-of-speech tags. The columns give information about the following:

- spelling
- spellcheck—whether the string was accepted as a correct word by the Aspell spell-checker
- alphabetical—whether the word contains only alphabetical characters
- nchar—number of characters in the string

SUBTLEX–PL frequency measures:

- freq—count of how many times the type appears in the subtitles
- capit.freq—count of how many times the type was capitalized
- cd—percentage of film subtitles in which the type appears
- cd.count—count of film subtitles in which the type appears
- dom.pos—most frequent part of speech assigned to the type

- dom.pos.freq—how many times this part of speech was assigned to the type
- dom.lemma.pos—dominant lemma[6] for the type
- dom.lemma.pos.freq—how many times this lemma was assigned to the type
- dom.lemma.pos.total.freq—total frequency of the most frequent lemma for the type (across all types)
- all.pos—list of all part-of-speech assignments for the type
- all.pos.freq—list of frequencies for all corresponding part-of-speech assignments in all.pos for the type
- all.lemma.pos—list of all lemma assignments for the type
- all.lemma.pos.freq—list of frequencies for corresponding lemmas in all.lemma.pos for the type
- all.lemma.pos.total.freq—total frequencies (across all types) of all corresponding lemmas in all.lemma.pos
- lg.freq—$\log_{10}$ of subtitle word frequency
- lg.mln.freq—$\log_{10}$ of subtitle word frequency per million
- zipf.freq—Zipf scale word frequency
- lg.cd—$\log_{10}$ of contextual diversity

Compound frequency measures:

- freq.sn.sum—sum of SUBTLEX-PL and BS–NCP word frequencies

[6] For practical reasons, we assume that lemma is equivalent to a concatenation of a base form of a word and an associated part of speech tag.

- zipf.freq.sn.sum—normalized (Zipf scale) sum of SUBTLEX-PL and BS–NCP word frequencies
- avg.zipf.freq.sn—averaged Zipf frequencies in SUBTLEX-PL and BS–NCP

The second file contains detailed information about lemma frequencies and particular forms for which this lemma was assigned. The columns in this file are the following:

- lemma—spelling of a base form of a lemma
- pos—part-of-speech tag assigned to a lemma
- spelling—word form assigned to a lemma
- freq—total frequency of a lemma or its inflected form
- cd.count—count of unique film subtitles in which the lemma or its inflected form appears
- cd—percentage of unique film subtitles in which the lemma or one of its inflected forms appears

Frequencies for word bigrams are included in a third file giving information about bigram frequency, contextual diversity, and all punctuation marks separating the words and their frequencies.

ACKNOWLEDGMENTS

This study was supported by an Odysseus grant awarded by the Government of Flanders to M.B. and a subsidy from the Foundation for Polish Science (FOCUS program) awarded to Z.W. We thank Jon Andoni Duñabeitia, Gregory Francis, and an anonymous reviewer for insightful comments on an earlier draft of the manuscript, Adam Przepiórkowski for providing access to the BS–NCP word

frequencies, and Jakub Szewczyk for his help with syllabification of Polish words.

REFERENCES

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual Diversity, Not Word Frequency, Determines Word-Naming and Lexical Decision Times. *Psychological Science*, *17*(9), 814–823. doi:10.1111/j.1467-9280.2006.01787.x

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. doi:10.1016/j.jml.2009.09.005

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General*, *133*(2), 283–316. doi:10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. Retrieved from http://link.springer.com/article/10.3758/BF03193014

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology, 3*(3), 296–322.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The Word Frequency Effect: A Review of Recent Developments and Implications for the Choice of Frequency Estimates in German. *Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie)*, *58*(5), 412–424. doi:10.1027/1618-3169/a000123

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. doi:10.3758/BRM.41.4.977

Brysbaert, M., & Diependaele, K. (2013). Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods*, *45*(2), 422–430. doi:10.3758/s13428-012-0270-5

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS One*, *5*(6), e10729. Retrieved from http://dx.plos.org/10.1371/journal.pone.0010729

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3), e57410. doi:10.1371/journal.pone.0057410

Cuetos Vega, F., González Nosti, M., Barbón Gutiérrez, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica: Revista de metodología y psicología experimental*, *32*(2), 133–143. Retrieved from http://dialnet.unirioja.es/servlet/articulo?codigo=3663992

Dimitropoulou, M., Duñabeitia, J. A., Avilés, A., Corral, J., & Carreiras, M. (2010). Subtitle-Based Word Frequencies as the Best Estimate of Reading Behavior: The Case of Greek. *Frontiers in Psychology*, *1*. doi:10.3389/fpsyg.2010.00218

Gale, W., & Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, *2*, 217–237. Retrieved from http://www.grsampson.net/AGtf.html

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Comput. Stat. Data Anal.*, *52*(12), 5186–5201. doi:10.1016/j.csda.2007.11.008

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. doi:10.3758/BRM.42.3.627

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. doi:10.3758/BRM.42.3.643

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice Effects in Large-Scale Visual Word Recognition Studies: A Lexical Decision Study on 14,000 Dutch Mono- and Disyllabic Words and Nonwords. *Frontiers in Psychology*, *1*. doi:10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2011). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. doi:10.3758/s13428-011-0118-4

Kilgarriff, A. (2006). BNC database and word frequency lists. Retrieved May 25, 2014, from http://www. kilgarriff.co.uk/bnc-readme.html

Korpus Języka Polskiego Wydawnictwa Naukowego PWN. Retrieved January 9, 2014, from http://korpus.pwn.pl/

Kurcz, I., Lewicki, A., Sambor, J., Szafran, K., & Woroniczak, J. (1990). *Słownik frekwencyjny poszczyzny współczesnej*. Kraków: Instytut Języka Polskiego PAN.

Lewis, M. P., Simons, G., & Fennig, C.D. (Eds.). (2013). *Ethnologue: Languages of the World, Seventeenth edition.* Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(04). doi:10.1017/S014271640707035X

Piasecki, M. (2007). Polish Tagger TaKIPI: Rule Based Construction and Optimisation. *Task Quarterly*, *11*(1–2), 151–167.

Przepiórkowski, A. (2012). *Narodowy Korpus Jezyka Polskiego: praca zbiorowa*. Warszawa: Wydawnictwo Naukowe PWN.

Przepiórkowski, A., & Instytut Podstaw Informatyki. (2004). *The IPI PAN corpus: preliminary version*. Warszawa: IPI PAN.

Schreuder, R., & Baayen, R. H. (1997). How Complex Simplex Words Can Be. *Journal of Memory and Language*, *37*(1), 118–139. doi:10.1006/jmla.1997.2510

Siyanova-Chanturia, A., Conklin, K., & van Heuven, W. J. B. (2011). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(3), 776–784. doi:10.1037/a0022531

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*(3), 271–295.

Woliński, M. (2006). Morfeusz — a Practical Tool for the Morphological Analysis of Polish. In M. Kłopotek, S. Wierzchoń, & K. Trojanowski (Eds.), *Intelligent Information Processing and Web Mining* (Vol. 35, pp. 511–520). Springer Berlin Heidelberg. Retrieved from doi:10.1007/3-540-33521-8_55

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *The*

Quarterly Journal of Experimental Psychology, *67*(6), 1176-1190. doi:10.1080/17470218.2013.850521

Vuong, Q. H. (1989). Likelihood Ratio Tests for Model Selection and non-nested Hypotheses. *Econometrica, 57*(2), 307–333.

Chapter 4. An exposure-based account of the changes in the word frequency effect

ABSTRACT

Although word frequency is usually measured using a logarithmic scale, the relationship between log-transformed frequencies and behavioral data is not completely linear but tends to flatten out for high frequency words. It is also well known that the size of the frequency effect changes depending on reader's proficiency. We consider whether statistical properties of a language sample (extreme distribution of word frequencies, underspecification of frequencies in the low frequency range) combined with a practice effect described by a power function can account for these findings. We demonstrate that these factors explain multiple phenomena observed in lexical research. We do so using corpus simulations and response time and accuracy measures collected in two massive word recognition experiments for English and Dutch with almost 1.5 million participants representing various demographic groups.

INTRODUCTION

Word frequencies in text corpora span an extremely wide range of values: a few very frequent words occur multiple times in one paragraph while others have minuscule probabilities of occurrence. Such extreme values are not typically observed in behavioral measures in standard psycholinguistic experiments and although word frequency is one of the strongest predictors of human performance in such tasks, that is the case only after transforming the frequencies. Dating back to the seminal study by Howes and Solomon (1951), the logarithmic transformation is most commonly applied in this context. While the logarithmic transformation is both convenient and easy to understand, it lacks a clear theoretical justification. However, the simple assumption that learning to recognize words is not fundamentally different from acquisition of other skills can lead to an alternative transformation. McCusker (1977; after Murray & Forster, 2004) and later Murray and Forster (2004) already tried to address this issue. They considered whether a power law (Newell & Rosenbloom, 1981) or an exponential function (Heathcote, Brown, & Mewhort, 2000), which are often considered to accurately describe the relationship between performance and practice, could also provide an adequate description of how frequency influences recognition time. They rejected this possibility by reasoning that this kind of asymptotic function would predict a diminishing word frequency effect with increased exposure to language, for example in older participants on the grounds that it was inconsistent with the available data. Before we revisit this prediction in greater depth, we review a few

methodological innovations and associated empirical findings that can shine some more light on this issue.

The first methodological innovation in lexical research is the development of megastudies, in which data for a large number of stimuli are collected and can be subsequently used to test various hypotheses (for a review see Keuleers and Balota, 2015). Importantly, because they provide behavioral measures for a large number of stimuli, these studies enable the application of regression analysis instead of factorial experiments. This approach is conductive to investigating various effects in more detail, including the functional relationship between word frequency and word processing efficiency. In the current paper, we will make use of megastudy data from participants covering a broad range of demographic characteristics.

The second important development is a more nuanced view on the use of text corpora in psycholinguistics. Firstly, currently available text corpora are much larger than the ones used for a long time in psycholinguistics. Secondly, several studies have shown that frequencies from certain types of textual materials, such as movie subtitles, are more adequate for use in psycholinguistics (e.g., Brysbaert & New, 2009; Keuleers, Brysbaert, & New, 2010). Finally, distributional properties of text corpora are also considered in the context of psycholinguistics (Kuperman & Van Dyke, 2013). This is interesting given that this combination can create a synergy between psycholinguistics and a broad body of knowledge accumulated in corpus linguistics. There is a reason to believe that corpus statistics should be looked at more carefully by psycholinguists because the statistical properties of text corpora can be assumed to also characterize the language samples on which the human linguistic

system is trained, potentially leaving its trace on how humans process language. For example, Blevins, Milin and Ramscar (2015) argue that regularity in language may be an effect of gaps in the paradigms of many words that are associated with Zipfian distribution of words.

In the current paper we make use of large text corpora, including corpora of movie subtitles, and, by conducting corpus based simulations, we investigate how statistical properties of language samples could affect human performance in psycholinguistic experiments.

Logarithmic functions have the mathematical property that the difference between the logarithms of two numbers remains constant when both of these numbers are multiplied by a third number. Therefore, when modeling differences in the processing characteristics of two words, it does not matter whether logarithms of relative or absolute frequencies are used. Because of this property, the difference between the log frequencies of two words does not change with increasing total exposure as long as the relative frequencies stay the same. After logarithmic transformation, the difference between one word and another word that is 10 times less probable is the same for a person who has experienced the first word 100 times and the second word 10 times as for a person who was experienced the first word 10 times and the second word once. As a result, the logarithmic function does not predict changes in the amount of the frequency effect with increased exposure (see also Figure 1).

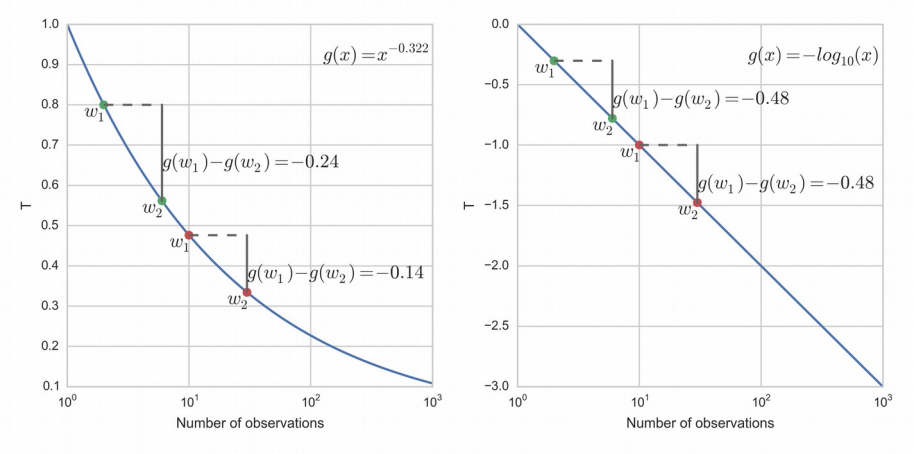

Figure 1. Predicted differences in time required to recognize word w_1 occurring with probability p_1 = 2e-6 and word w_2 occurring with probability p_2 = 6e-6 after 1 million learning trials (green) and 5 million (red) learning trials according to a power function (left) and a log function (right). The difference decreases asymptotically for any pair of probabilities p_1 and p_2 for a power function but remains constant for a log function.

Interestingly, the empirical findings, which are largely based on the methodological innovations listed above, seem to suggest that the relationship between behavioral measures and word frequencies is not exactly logarithmic. It is quite clear that the relationship between log transformed word frequency and response times in lexical decision (Keuleers, Lacey, Rastle, Brysbaert, 2012; Keuleers, Diependaele, & Brysbaert, 2010; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004) or eye-tracking experiments (Cop, Keuleers, Drieghe, & Duyck, 2015; Kuperman & Van Dyke, 2013) is not completely linear but tends to flatten out for the high frequency words. In fact, the word frequency effect is not distributed equally across the entire frequency range but has been observed to concentrate in the range between 1 per

million and 10 per million. Van Heuven, Mandera, Keuleers and Brysbaert (2014) even proposed that a new scale which is centered in the middle of its typical range, called a Zipf scale, should be used to measure word frequencies.[1]

Secondly, it has been regularly observed that the word frequency effect becomes less steep for more proficient participants, both when comparing behavioral measures collected from non-native and native speakers of a language (Van Wijnendaele & Brysbaert, 2002; Gollan, Montoya, Cera, & Sandoval, 2008; Duyck, Vanderelst, Desmet, & Hartsuiker, 2008; Gollan et al., 2011; Whitford & Titone, 2012; Lemhöfer et al., 2008) and within groups of native and non-native speakers as a function of their language proficiency (Diependaele, Lemhöfer, & Brysbaert, 2012). Increased proficiency can be considered to be associated with more experience with a language so these results are compatible with other studies that looked at the size of the frequency effect as a function of the amount of exposure to print (Chateau and Jared, 2000) or proficiency (Pugh et al., 2008, Shaywitz et al., 2003).

Although the empirical relationship between the size of the word frequency effect and language proficiency is rather uncontroversial, its interpretation is much less clear. Diependaele and colleagues (2012) contrasted two potential explanation of this effect in bilinguals, a structural one, according to which competition between

[1] The Zipf scale is a logarithmic measure of the number of occurrences per billion words with Laplace smoothing. The scale was proposed as a more convenient scale on which word frequencies may be measured. In order to reflect the nature of the frequency effect, it is a logarithmic scale (like the decibel scale of sound intensity), but, in contrast to the logarithm of frequency per million words, it does not result in negative values for corpora of up to 1 billion words.

two languages can lead to an increased word frequency effect and one based purely on within language characteristics. They found that the size of the frequency x skill interaction can be fully explained based on within-language factors, clearly favoring the second interpretation. Kuperman and Van Dyke (2013) proposed that the interaction between skill and word frequency effect may be an artifact attributed to overestimation of the word frequencies in the low frequency range when applying large text corpora to explain the behavior of participants who have much less language experience. They also show that the interaction between skill and proficiency is removed when using subjective word frequencies instead of word frequencies derived from text corpora.

If one considers that reading proficiency is likely to be strongly associated with the total amount of exposure to language (even if this is not the only influencing factor), arguments against the power function as describing the functional relationship between frequency and behavioral measures becomes less obvious. A critical piece of evidence on which Murray and Forster (2004) rejected the power function as a potential candidate for such a relationship was based on empirical studies showing equal or stronger frequency effect in older participants, who should have more language exposure and should thus show a smaller frequency effect, than younger participants (Tainturier, Tremblay, & Lecours, 1989; Balota & Ferraro, 1996; Spieler & Balota, 2000). However, given the direction of the skill x proficiency interaction described above, the results of the studies which looked at the relationship between age of participants and the size of the word frequency effect are puzzling. Finding a larger effect in older participants would be paradoxical, considering that there is

good evidence that older participants have a larger vocabulary size (O'Dowd, 1984; see also Keuleers, Stevens, Mandera & Brysbaert, 2015 analyses based on a part of the dataset used also in the current paper) and that vocabulary size was used as a measure of language proficiency in many studies that found a frequency x skill interaction with the opposite pattern (Diependaele et al., 2012). One of the reasons why this paradoxical effect of age may have been observed, could be that in the studies investigating this topic, typically a group of very young adults (typically university students) is contrasted with a group of much older adults (typically around 70 years old). Instead of focusing on the extreme groups, it would be more beneficial to look at how the word frequency effect changes across the entire lifespan rather than focus on the two groups of extreme ages. The question of how the word frequency effect changes across lifespan is also interesting in the context of the discussion about the existence of an age-related cognitive decline. Because it has been argued that some effects associated with aging may be a simple consequence of learning (Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014), it would be interesting to know whether there is a consistent pattern of changes in the word frequency effect associated with systematic exposure to linguistic stimuli.

Although not directly related to the word frequency effect, it is interesting that, based on the analysis of a large dataset of lexical responses (Keuleers, Stevens, Mandera, & Brysbaert, 2015), we observed that the pattern of the increase in vocabulary size with increased age is remarkably similar to the vocabulary growth curve observed in text corpora. This pattern is approximated by Herdan's law (Herdan, 1960), which states that the probability of encountering a

new word type decreases with the number of encountered word tokens, which results in vocabulary size increasing at an ever slower pace.

This is an interesting observation, also in the context of the current discussion, because it suggests that, given a large enough sample of participants and linguistic stimuli, it may be possible to observe behavioral patterns reflecting the properties of the language samples experienced by the participants. In the current paper, we re-examine the predictions made by the power function in describing the word frequency effect's dependence on the size of the language input that different groups of participants have experienced. We first make theoretical predictions that are the consequence of applying power functions to samples of language. Importantly, we do so in the context of language statistics. We support these considerations with corpus simulations. Next, we describe two web based word knowledge experiments in Dutch and English, with a total of nearly 1.5 million participants. We evaluate the quality of the response times collected in these experiments, and, finally, show that the patterns expected purely based on applying a power function to language samples, can be also observed in the word recognition data. Importantly, in our analyses we do not rely on fitting different non-linear functions (such as the logarithm and the power function) to empirical data to distinguish which one best describes the dataset, as such approach can easily lead to spurious findings (Clauset, Shalizi, & Newman, 2009). Instead, we test predictions that the power functionand logarithmic function make regarding the changes in the frequency effect in groups with different degrees of linguistic exposure.

LANGUAGE STATISTICS AND THE POWER FUNCTION

It is known that a practice curve is asymptotic – the time to perform a task generally decreases with practice but for each subsequent repetition the improvement becomes smaller. It was proposed that this relationship can accurately be described by a power function [2](Newell & Rosenbloom, 1981):

$$T = BN^{-\alpha} \qquad \text{(Eq 1.)}$$

where B and alpha are parameters and N is the trial number (repetition of a task).

One obvious consequence of using this function to model word frequencies is that it can account for the nonlinearity in the word frequency effect that is observed when word frequencies are measured on the logarithmic scale. The power function first leads to a faster decrease in reaction times than the logarithmic function but leads to a slower decrease later on. This results in a pattern that compensates for the flattening out of the response times for high frequency words and predicts a stronger frequency effect in the lower frequency ranges. Secondly, the power function makes concrete predictions for different words not only based on their relative frequency in a language but also depending on the total amount of experience that a person has. Thirdly, its predictions can be easily tested by combining the power

[2] It has been argued that an exponential function may better approximate practice effect when data from individual participants are considered (Heathcote, Brown, & Mewhort, 2000) and that the observed power function between practice and performance may be a consequence of averaging data from individual participants. In this paper, we work with data aggregated over participants so we consider the power function to be a sufficient description. See also the comments in the Discussion.

function with corpus-based simulations if we assume that differences in the amount of exposure to language in human participants can be modeled in terms of the size of the language sample that the person has been exposed to.

If we consider the recognition of each individual word to be a task that has to be mastered, and if we know the probabilities with which words occur in a language, we can easily modify Eq 2. to describe the functional relationship between the sample of a language that the person has experienced and the recognition time of a given word:

$$T = B\left(p_w S\right)^{-\alpha} \qquad \text{(Eq 2.)}$$

where p_w is the probability of the word and S is the size of the language sample.

The word frequency effect can be rephrased as a difference in response time to two words, p_1 and p_2, which can be expressed as:

$$T_2 - T_1 = B\left(p_2 S\right)^{-\alpha} - B\left(p_1 S\right)^{-\alpha} \qquad \text{(Eq. 3)}$$

For any pair $p_1 > p_2$, this difference approaches zero as the sample size approaches infinity (see also Figure 1.), so Murray and Forster (2014) correctly recognized that the entire word frequency effect should eventually disappear (for a finite vocabulary size).

However, words are not experienced as independent phenomena – they are always part of a larger sample. Therefore, if we are considering difference associated with the recognition of two words occurring with a given frequency in a language, we should consider them as parts of a frequency distribution and not in isolation.

The distribution of word frequencies was seminally described by Zipf (1949) and is associated with a number of properties.

Firstly, the consequence of the Zipfian distribution of word frequencies is that probabilities of words in a language vary across multiple orders of magnitude: there are some words with very high probabilities but most words are concentrated in the extremely low part of the frequency distribution. What are the consequences of this fact for the differences based on the power function? Eq 3. describing the difference in required effort to recognize two words, can be simplified to:

$$T_2 - T_1 = B\left(p_2^{-\alpha} - p_1^{-\alpha}\right) S^{-\alpha} \qquad \text{(Eq 4.)}$$

So the difference in processing time is proportional to $p_1^{-\alpha} - p_2^{-\alpha}$. If we assume that we use a $\log_{10}$ scale, as is usually done with word frequencies, then each unit on a log scale is associated with an order of magnitude change in word probabilities. In other words, $p_1 = 10p_2$, and:

$$p_2^{-\alpha} - p_1^{-\alpha} = \left(1 - 10^{-\alpha}\right) p_1^{-\alpha} \qquad \text{(Eq 5.).}$$

From this follows that the difference depends on the p_1 in such a way that the extreme differences in p_1 cause large differences in the size of the predicted frequency effect. At the same time the speed at which this difference is changing is equal to:

$$\frac{dT_2 - T_1}{dS} = -\alpha B\left(p_2^{-\alpha} - p_1^{-\alpha}\right) S^{(-\alpha-1)} \qquad \text{(Eq 6.)}$$

In other words both the speed at which the difference in the response times for two words decreases and their absolute values are

larger in the low frequency range. It is also clear that the difference will reach very low values for high frequency words very quickly, but that this is not the case in the low frequency range (see Figure 2.).

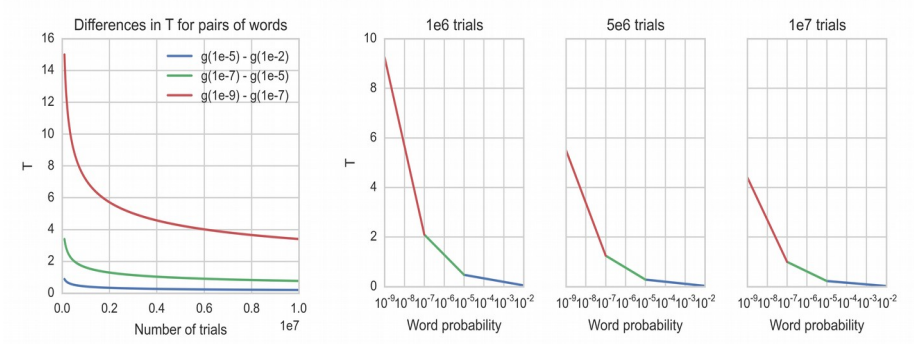

Figure 2. The leftmost panel shows predicted differences in time required to process two low frequency words (red) occurring with probability 1e-9 (Zipf value 0) and 1e-7 (Zipf value 2), medium frequency words (green) occurring with probability 1e-7 (Zipf value 2) and 1e-5 (Zipf value 4) and high frequency words (blue) occurring with probability 1e-5 (Zipf value 4) and 1e-2 (Zipf value 6) as a function of total amount of experience with language. The remaining three panels show slopes of predicted lines connecting these pairs of words after 1 million trials, 5 million trials and 10 million trials. Similarly, to what is observed in behavioral data the frequency effect decreases. The effect rapidly becomes very small for the high frequency words. The difference between the medium and high frequency pairs of words decreases fast but is still robust even after a substantial number of learning trials.

In general, this is in line with the observation that the word frequency effect is stronger for low frequency words than for high frequency words regardless of the proficiency level of a participant. At the same time, the difference between the more and less proficient participants, corresponding to larger and smaller language samples that these participants have experienced respectively, is predicted to be

larger for the low frequency words than for the high frequency words as observed in empirical studies (e.g. Diependaele et al., 2013).

Another property of the Zipfian distribution is that it is very difficult to obtain precise frequency estimates for the low frequency words. In the above simulations we assumed that the frequency values constitute a fully continuous variable and can take any real value. In reality, observed frequencies can only take integer values. In consequence, it is known that a very large proportion of the words is either unobserved in a language sample of any size (have frequency 0), or occurs only once (frequency 1). This is true for any sample size if we consider a theoretical Zipfian distribution, but also in practice if we consider a text corpus of a realistic size and a finite vocabulary (Baayen, 2001).

As a consequence, differences in frequency between high frequency words become stable even in very small samples, but differences between low frequency words remain singular (see Figure 3) and a much larger sample size is necessary to differentiate between frequencies of the low-frequency words.

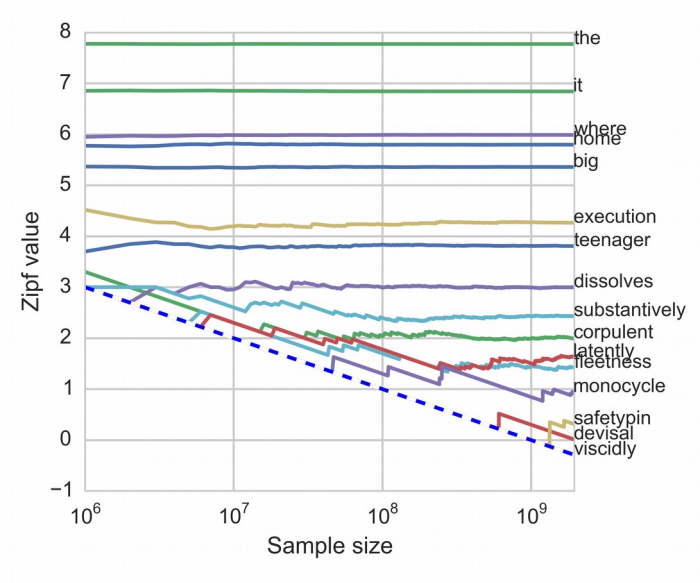

Figure 3. Relative frequency estimates for a set of words distributed over the entire range of frequencies in the UKWAC corpus, shown as a function of the sample size taken from the corpus. Relative frequencies of high frequency words are stable even in small sample sizes. In the case of the low frequency words, even in relatively large sample size many words have singular frequencies or noisy frequency estimates. The blue dashed line shows Zipf value associated with words that have frequency 0 in the respective sample sizes.

To understand what this implies for the frequency effect, let us consider a hypothetical language user who has experienced only 100 words in her entire life. Assuming perfect retention, a power function predicts such a person to have a very strong frequency effect since all the words will be relatively far from the asymptote. On the other hand, roughly 10 of the words that she experienced will be just the most frequent words, a few others will have frequencies 2 to 4, and the remaining tens of words will have frequency 1. The vast majority of the words will not be observed at all. As a result, for the

extremely high frequency words, which would likely have been the ones with frequencies larger than 1 in the sample, we can expect to see a very strong frequency effect. Although the effect should be even stronger for the low frequency words based on the predictions of the power function, it is impossible to observe it as there is no variability for the low frequency words. As we keep increasing the sample size, the frequency effect will become increasingly observable in the very low frequency range, even though at the same time, due to the nature of the power function, practice would decrease the underlying effect. Necessarily, for any frequency range and at any point in time one of these two opposite tendencies will dominate: as we increase the size of the sample, the underlying difference will decrease due to the asymptotic nature of the practice function, but in the low frequency range the effect is not observable in the beginning due to underspecification of the frequencies. Any increase in the sample size will increase the observable effect for the low frequency words.

Importantly, the inaccuracies in the low frequency range and the size of the frequency effect for the higher frequency range predicted by the power function are not independent phenomena but are tied together because in any language sample the absolute frequencies of the high frequency words are necessarily tied to the amount of the variability and the quality of the estimates for low frequency words and vice versa.

It is reasonable to ask how the sampling inaccuracies would be expressed in the behavioral data. For accuracy, it is quite clear that one should not be able to recognize a word that one has never seen and that the probability of recognizing a word should be an increasing function of the experience with this word until it reaches a ceiling

effect (which does not need to be equivalent to the asymptote of the learning effect). For reaction times, because typically only the correct responses are usually considered in the response times, first of all we should expect to have fewer observations for the low frequency words. It has also been reported (Diependaele, Brysbaert & Neri, 2012) that both for accuracy and response times participants are much more probable to respond randomly (the responses are more noisy) for the low frequency words. All in all we would expect to find less responses in the low frequency range and more random responses, which should result in lack of a reliable effect for such words.

A CORPUS-BASED SIMULATION OF THE SIZE OF THE FREQUENCY EFFECT

In order to evaluate how these predictions would materialize in language samples of varying size–which could be assumed to reflect varying exposure to language in participants with different age, proficiency, educational level, etc.– we conducted a corpus simulation based on UKWAC, a corpus of about 2 billion words resulting from a web crawl (Ferraresi, Zanchetta, Baroni, & Bernardini, 2008). First, we selected 150 random sample sizes in the range between 2 million and 200 million words. For each of the sample sizes, we selected a random starting point in the UKWAC corpus and calculated word frequencies based on the portion of the corpus from the starting point until the desired sample size was reached.

Next, we looked at the speed with which the correlations between the frequency estimates derived from different sample sizes and the word frequencies calculated based on the entire corpus increased in different frequency ranges. In order to look at individual

frequency bands we split the full frequency range into three parts: the high frequency range including all words with Zipf values higher than 4, the medium frequency range including all words with Zipf values between 2 and 4, and the low frequency range including all words with Zipf values below 2. The frequency ranges were defined in this way because they divide the full frequency range more or less equally and roughly correspond to different parts of the frequency curve reported in lexical decision tasks (Keuleers, Diependaele, & Brysbaert, 2010; Keuleers, Lacey, Rastle, & Brysbaert, 2011; Van Heuven et al., 2014).

The pattern of obtained correlations can be seen in Figure 4. As expected based on the properties of the Zipfian distribution, in the highest frequency range the estimates were almost perfect even in the smallest samples. In the medium frequency range, the increase in correlations was smaller but also reached a very high level quite fast. In the lowest frequency range, however, the correlations increased rather slowly and even for the sample including 200 million words did not reach 0.9.

Next, in order to simulate the practice effect we applied a power function to the frequencies in each of the corpus samples. We used an arbitrary exponent equal to -0.322 which corresponds to a 80% learning rate (the time required to perform a task drops to 80% of the value with each doubling of the number of learning trials). The choice of this exponent was arbitrary, but the purpose of this simulation was to demonstrate a general consequence of the practice combined with different language sample sizes, and a similar pattern can be observed with other values of this parameter.

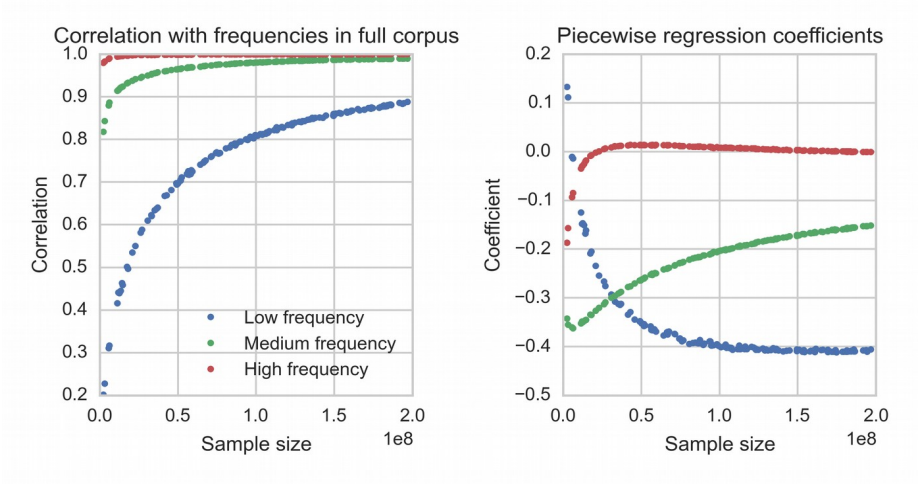

Figure 4. Correlations between log 10 of the word frequency estimates in the full UKWAC corpus and samples of varying size in different frequency ranges (left panel). The correlation between the sample frequencies and the full corpus frequencies increases rapidly in the high and medium frequency range, but for the low frequency the increase is much slower. The right panel shows slopes fitted using a piecewise regression in the three different frequency ranges, after applying a power function to word frequencies in the different sample sizes. For the high frequency words, the frequency effect is present only for extremely small sample sizes. In the medium frequency range the slope slowly becomes less steep. For the low frequency words, initially the slope becomes steeper with increasing sample sizes due to improvement in correlation between the word frequencies derived from a sample and the word frequencies in the full corpus. Note that the frequency effect results in negative slopes and that an absence of a frequency effect corresponds to a slope of 0.

Finally, we conducted a regression analysis to quantify the amount of the word frequency effect in each of the frequency ranges. In contrast to the theoretical exposition presented earlier, in which we considered only differences in predicted response times between pairs of words, in the case of the language samples we had to deal with words being dispersed across the entire range of word frequencies. In

order to quantify the amount of the word frequency effect, while simultaneously taking into account all words, we applied a piecewise regression analysis, with slopes free to change in each of the frequency ranges (defined in the same way as in the correlation analysis). We used Zipf values derived from the full UKWAC corpus as an independent variable and the response times predicted by applying the power function to word frequencies in each sample size as a dependent variable. The piecewise regression technique allowed us to model changes in the shape of the word frequency effect in the different frequency ranges.[3]

The results of the corpus simulations can be seen in Figure 4. As expected, in the highest frequency range the coefficient was approaching 0 even for very small samples (corresponding to low-proficient participants) as the absolute frequencies of the high frequency words increased rapidly and the effect diminished due to the properties of the power function. As could be predicted based on the mathematical derivation presented above, the frequency effect in the medium frequency range was much more robust and was decreasing slowly with increasing sample size. In the lowest frequency range a reversed pattern was observed: increasing sample size led to an increased frequency effect. This effect may seem paradoxical but becomes quite clear if you consider the slow increase in the correlation between sample and full-corpus word frequency estimates in that frequency range as well as the fact that the power function would predict a strong frequency effect in that range.

[3] We confirmed the analyses conducted using the piecewise regression, by fitting a completely independent regression lines in each of the frequency bands. However, this did not change the qualitative patterns of results, so these analyses are not reported in the paper.

In summary, this simulation fully confirms the theoretical predictions presented in the previous section. Increased sample size, which we consider to be equivalent to increased language exposure, when combined with a power function and properties of the Zipfian distribution leads to a rapid decrease in the word frequency effect for the high frequency words, a slower but systematic decrease in the slope for medium frequency words, but can lead to the reversed pattern in the case of the low frequency words, where increasing accuracy of sample estimates outweighs the decreases predicted by the power function. Assuming that language proficiency and exposure are equivalent to increased experience with words, we can expect to see the same patterns for more proficient and experienced participants with increasing sample size from a text corpus.

Although these simulations have shown the predictions regarding different language sample sizes, in psycholinguistic experiments we often deal with averages of measures collected from multiple participants. Because of that it should be considered whether the line of thought which led us to say that frequency estimates in the low frequency range are unreliable also applies to frequencies averaged across different samples. In order to answer this question we conducted another corpus simulation in which we drew 100 samples of 1 million tokens and 100 samples of 3 million words and looked at whether the correlation between the average of the $\log_{10}$ frequencies in smaller sample sizes can be used to approximate a larger sample size (see Figure 5.)

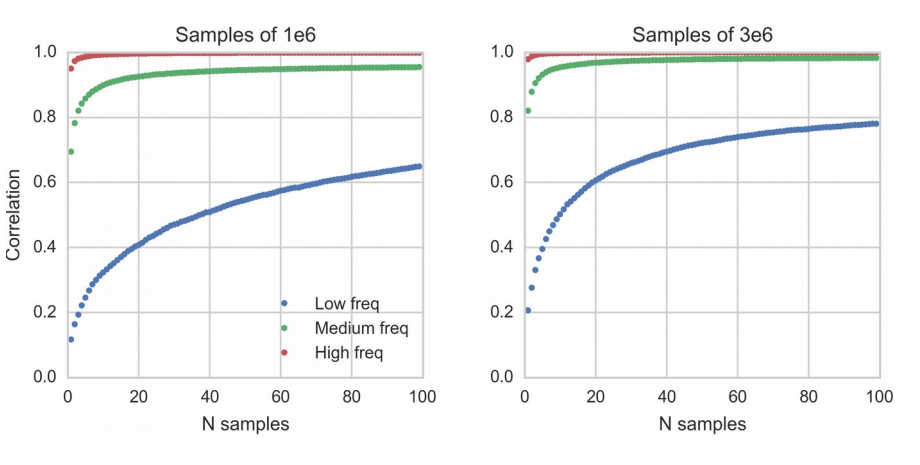

Figure 5. Effect of averaging sample sizes of 1 million (left panel) and 3 million words (right panel) on the correlations between the log 10 of the averaged frequencies and log 10 of word frequencies in the full corpus. For the high and medium frequency words both sample sizes give an almost perfect correlations with the estimates derived from the full corpus after averaging frequencies from a relatively small number of samples. For the low frequency words, correlations increase more slowly. The correlations increase faster when averaging larger sample sizes than when averaging smaller sample sizes.

We see that an average of smaller samples gives lower correlations with the full corpus than the same number of larger samples. Based on this we can conclude that our use of single samples instead of averaging a larger number of smaller samples in the previous simulation is a valid approach.

MEGASTUDIES

The question about the shape of the word frequency effect is a question about the shape of the relationship between two variables. In this case it is important to consider both behavioral measures and

word frequency as continuous variables. The benefits of this approach over factorial designs has been discussed in the literature (Balota, Yap, Hutchison, & Cortese, 2012; Balota & Keuleers, 2015; Baayen, 2010). The regression designs work particularly well if they are combined with the megastudy approach (Balota et al., 2004; Keuleers et al., 2010; Keuleers et al., 2011; Brysbaert, Stevens, Mandera, & Keuleers, 2015), in which data for a large number of stimuli are collected.

Despite the fact that the megastudy approach has proven to be successful in many ways, the creation of this kind of datasets is resource consuming because of the large number of words in a language that are potential stimuli. Moreover, although megastudies are so comprehensive and constitute an almost complete snapshot of a language in terms of the range of stimuli that they cover, they are very limited when it comes to whose language they represent. They are based almost exclusively on the language of fairly young participants, native speakers of a language, typically undergraduate students.

To test the evolution of the word frequency across groups with varying the amount of exposure, such as participants with different educational level and age, we would ideally have megastudy data for each of the groups.

The problem with collecting megastudy data while taking into account different demographic groups is that there is a multiplicative relationship between the number of observations required and the number of levels that we want to consider, which is a problem if we consider the high number of trials required to cover the wide range of linguistic stimuli.

COLLECTION OF REACTION TIMES

Due to practical limitations, collecting data for all words from a wide range of demographic groups in traditional laboratory-based settings would be very challenging. So, we decided to investigate the associated changes in the shape of the word frequency effect using reaction times collected in a web based word knowledge experiment that we recently conducted. Although, collection of human ratings through various Internet-based platforms is now a well-established method (Brysbaert, Warriner, & Kuperman, 2014; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Warriner, Kuperman, & Brysbaert, 2013), this kind of collection of other types of behavioral data such as accuracies and, in particular, the RTs is a topic of increasing importance. It has been shown that this method of data collection can be useful, although caution is required with respect to some experimental procedures and experimental designs. For example, Crump, McDonnell and Gureckis (2013) used Amazon Mechanical Turk to replicate some classical psychological effects. They attempted a replication of the Stroop, switching, flanker, Simon effect, attentional blink, subliminal priming, and category learning task. They managed to replicate most of these effects, including effects as small as 20-ms but failed to replicate the effect of masked priming, likely due to technical issues with the control of presentation time, and category learning, likely due to lack of sufficient motivation in their participants. More recently, Reimers and Stewart (2015) conducted a detailed evaluation of the presentation and measurement accuracy of different systems that can be used to conduct Web-based experiments. They found that within-system variability is rather small. However, the between-system variabilities can be substantial. This

finding has consequences for the types of experimental designs that can be used with on-line collection of reaction times. Because of the low variability within systems, collection of a slightly larger number (10% according to Reimers and Stewart, 2015) of observations is sufficient to compensate for the increased amount of noise. The between system variability, however, can be more problematic in some cases. Especially, when between-subjects comparisons are made based on their absolute response times, spurious correlations can be found. For instance, if older participants are more probable to use Web browsers which give slower reaction time measurements, this can result in spurious correlations between age and an absolute reaction time. If we consider, however, within-subject effects such as a Stroop effect, even if the measurement in some browsers is systematically longer or shorter, it is the same for both conditions, so the effect should be the same.

In summary, we need to be cautious when dealing with experimental procedures that require an extremely precise timing in presentation of stimuli (such as masked priming; Crump et al., 2013) and where we intend to make between-subject comparisons of absolute response times (Reimers, & Stewart, 2015). The type of analyses conducted in this paper is largely unaffected by these problems: the word knowledge task does not depend on extremely accurate presentation times and our analyses compare differences between response times for different stimuli (the word frequency effect) rather than absolute reaction times. In addition, there is no reason to believe that some systems will systematically overestimate reaction times selectively for low frequency words or high frequency words.

To further control for the influence of differences between platforms, in all analyses we also use standardized reaction times per participant in addition to the raw reaction times. This procedure removes potential influences associated with using various platforms as we compute them for individual sessions which use the single platform. We also correlate our values with existing databases of reaction times.

METHOD

PROCEDURE

We conducted two word knowledge studies – in Dutch (http://woordentest.ugent.be) and in English (http://vocabulary.ugent.be). The analyses reported in the current paper are based on data collected over a period from March 2013 to December 2013 for Dutch and January 2014 until August 2015 for English.

In total 54,333 words and 21,748 pseudowords were used in the Dutch test and 61,856 words and 329,851 pseudowords in the English test. The pseudowords were created using Wuggy (Keuleers & Brysbaert, 2010).

Both studies were based on the principle that the experiment should be short and that nothing is obligatory. After accessing the website, participants were presented with an instruction asking them to indicate for each letter sequence that would be presented on the screen whether they knew the word. They were informed that some letter sequences were made-up words and that their final score would be penalized if they responded 'yes' to these pseudowords.

Because the test was designed to work on computers but also on other devices such as smartphones and tablets, the instruction was tailored separately for devices with a physical keyboard and for devices with a touchscreen. For the keyboard devices the instruction indicated that the buttons 'j' and 'f' should be used to give 'word' and 'non-word' responses respectively. On touchscreen devices, two buttons with 'yes' and 'no' labels (or Dutch translations of these words) were shown on the screen during the experiment and the responses were given by touching these buttons (see Keuleers et al, 2015, for more information).

After being presented with the instructions, participants were asked to give answers to a set of questions regarding their demographic background. In the Dutch test, questions regarding age, gender, location, education level, mother tongue, number of other languages known, best other language, level of knowledge of the other language, and handedness were asked. In the English test, a similar set of questions was asked with a difference that instead of asking for the level of the best other language we asked for the level of English. Answering any or all of these questions was not required to proceed to the word knowledge part of the experiment.

Next, participants were presented with 100 items, which included 70 words and 30 pseudowords. For the Dutch test, sets of 100 items were collected into fixed lists before the start of the experiment. As a result, each word and pseudoword was always presented within the same set of words and pseudowords, although the order of presentation within each list was randomized. In the English test, a new list was created by randomly sampling 70 words and 30 pseudowords every 5 seconds and this set was presented to all

participants starting the test in this 5-second time window. As a result most participants were presented with a unique set of stimuli. There was no time limit to give a response.

After completing the test, participants were presented with their score, calculated as a percentage of correctly recognized words minus the percentage of incorrectly accepted pseudowords. Participants were allowed to do the test multiple times. In the part of the experimental procedure in which participants were asked to fill-in the profile information, the answers to the questions given previously were shown. If nothing was changed in the profile information, we considered all experimental sessions completed without making a change in the profile as representing the same participant.

RESULTS

ENGLISH

In total, for the English test we collected about 89 million responses (words: mean = 1007, *SD* = 159; pseudowords: mean = 80, *SD* = 31) from 890 thousand experimental sessions. We performed basic cleaning of the full dataset in order to limit the amount of noise. We considered only responses from the 3 first sessions associated with each profile and considered only the responses from the 10th and subsequent responses given in the test. Trials 1-9 were considered as training trials although they were not explicitly specified as such in the instruction. We also removed all trials with responses longer than 8000 ms and subsequently removed exceedingly fast and slow responses using an adjusted boxplot method (Hubert & Vandervieren,

2008) calculated separately for words and pseudowords in each individual session.

After applying this cleaning procedure, 70 million responses (words: mean = 800, *SD* = 130; pseudowords: mean = 63, *SD* = 25) from 837 thousand sessions remained in the dataset. 299 thousand of these sessions were collected using devices with touchscreen and 538 thousand from keyboard devices. After applying the cleaning procedure, we calculated standardized reaction times (zRT) based on correct responses, separately for words and pseudowords in each experimental session.

Overall accuracy in the cleaned dataset was 73.0% for words and 91.4% for pseudowords. Average response times were equal to 1265 ms (*SD* = 859) for words and 1466 ms (*SD* = 925) for pseudowords.

DUTCH

In total, in the Dutch test we collected about 60 million responses (words: mean = 796, *SD* = 115; pseudowords: mean = 777, *SD* = 137) from 600 thousand experimental sessions. The same cleaning procedures as for English were applied. After cleaning, about 43 million responses (words: mean = 572, *SD* = 88; pseudowords: mean = 552, *SD* = 102) from 513 thousand sessions remained in the dataset. 139 thousand of these sessions were collected using devices with touchscreen and 373 thousand from keyboard devices.

Overall accuracy in the cleaned dataset was 83.8% for words and 87.2% for pseudowords. Average response time for correct trials was 1270 ms (*SD* = 778) for words and 1797 ms (*SD* = 1034) for pseudowords.

QUALITY OF THE COLLECTED REACTION TIMES

Reliability

First, in order to evaluate the quality of the collected reaction times we calculated split-half reliabilities of reaction time estimates for words and pseudowords. The two halves were created by randomly selecting half of the total number of experimental sessions, calculating the statistics for each of the stimuli, calculating the correlations between the measures derived from the two halves and applying a Spearman-Brown correction (Brown, 1910; Spearman, 1910).

The general reliability of the reaction times collected for words was almost perfect. It was equal to 0.98 for raw RTs and 0.99 for standardized RTs in the English test and to 0.97 for RTs and 0.99 for the standardized RTs in the Dutch test. For pseudowords, the reliabilities in the English test were equal to 0.77 for raw RTs and 0.88 for standardized RTs. In Dutch they were equal to 0.96 and 0.98 respectively.

Next, we considered the reliability of different subsets of the full dataset. We calculated reliabilities for various demographic groups by selecting only sessions collected from participants from that group and then following the same procedure as in the case of the full dataset.

Table 1. Information about individual subgroups in the Dutch word knowledge experiment.

		Words				Pseudowords			
				Reliability				Reliability	
	N sessions	Accuracy	Mean RT	RT	zRT	Accuracy	Mean RT	RT	zRT
age group (native speakers)									
0 - 9	7546	.82	1287.58	.26	.52	.84	1783.44	.21	.46
10 - 17	24254	.74	1410.18	.45	.65	.78	1794.83	.45	.64
18 - 23	60774	.79	1249.82	.75	.86	.84	1621.03	.74	.87
24 - 29	58257	.82	1299.77	.78	.88	.86	1725.63	.74	.87
30 - 39	77651	.84	1353.79	.83	.91	.87	1838.20	.80	.90
40 - 49	74029	.85	1341.80	.84	.92	.88	1846.15	.80	.91
50 - 59	72057	.87	1326.61	.84	.93	.89	1851.00	.80	.91
>= 60	60890	.88	1350.00	.82	.92	.89	1913.26	.78	.90
country (native speakers)									
Belgium	180248	.83	1316.86	.92	.96	.86	1822.05	.90	.95
Netherlands	239673	.84	1341.54	.94	.97	.88	1792.47	.92	.97
education (native speakers)									
No	1684	.77	1491.19	.04	.23	.75	2072.01	.11	.26
Primary	8462	.77	1524.85	.25	.48	.77	2083.20	.19	.40
Secondary	120902	.83	1354.80	.88	.94	.84	1857.36	.85	.93
Bachelor	156724	.84	1311.73	.91	.95	.87	1778.39	.90	.95
Master	143571	.85	1306.40	.91	.95	.90	1773.16	.89	.95
gender (native speakers)									
Female	229574	.83	1308.32	.93	.97	.87	1760.64	.92	.97
Male	202669	.84	1348.02	.93	.96	.87	1856.03	.92	.96
handedness (native speakers)									
Right-handed	375104	.84	1325.68	.96	.98	.87	1804.69	.95	.98
Left-handed	58573	.84	1327.30	.78	.89	.87	1802.97	.74	.87
best foreign language (native speakers)									
English	323897	.84	1316.89	.95	.98	.88	1778.26	.95	.98
French	46331	.85	1336.24	.78	.88	.87	1885.63	.72	.85
German	24307	.86	1342.56	.64	.81	.88	1854.46	.55	.76
level of best foreign language (native speakers)									
I know a few words.	3853	.77	1530.92	.16	.37	.77	2081.02	.12	.29
I can have a simple conversation.	42800	.83	1388.81	.73	.86	.85	1901.84	.69	.84
I can read a simple book.	55167	.83	1363.80	.78	.89	.87	1843.55	.73	.87
It is my second mother tongue.	31287	.84	1293.82	.67	.81	.85	1800.93	.59	.77
I speak and read the language fluently.	296317	.84	1310.03	.95	.98	.88	1779.21	.94	.97
number of foreign languages (native speakers; at least 1000 sessions)									
0	2166	.80	1475.63	.13	.35	.81	2104.97	.17	.28
1	43116	.81	1390.48	.70	.85	.84	1864.49	.66	.83
2	145184	.83	1343.97	.90	.95	.86	1820.54	.89	.95
3	167579	.84	1313.91	.91	.96	.88	1790.41	.90	.96
4	54552	.85	1285.83	.79	.89	.88	1767.34	.73	.87
5	14455	.86	1264.18	.52	.70	.88	1753.39	.39	.63
6	3760	.87	1224.41	.24	.43	.89	1734.59	.16	.36
native language									
Arabisch	1255	.75	1487.32	.19	.25	.72	2001.45	.21	.23
Duits	2226	.81	1457.21	.18	.31	.79	2076.32	.15	.36
Engels	2172	.79	1455.41	.15	.31	.79	2001.94	.12	.28
Frans	3896	.79	1453.57	.21	.36	.79	1996.26	.24	.37
Fries	2692	.84	1252.62	.20	.40	.87	1720.48	.15	.32
Nederlands	435458	.84	1325.94	.96	.98	.87	1804.55	.96	.98
Turks	1284	.75	1418.26	.14	.25	.72	1924.56	.14	.25

Table 2. Information about data collected from different subgroups in the English word knowledge experiment.

		Words				Pseudowords			
				Reliability				Reliability	
	N sessions	Accuracy	Mean RT	RT	zRT	Accuracy	Mean RT	RT	zRT
age group (native speakers)									
10 - 17	28693	.69	1174.64	.46	.67	.88	1230.93	.09	.19
18 - 23	69646	.74	1211.80	.76	.86	.91	1306.57	.20	.35
24 - 29	80071	.77	1265.63	.83	.90	.92	1414.14	.25	.42
30 - 39	103113	.79	1295.42	.88	.93	.92	1490.87	.32	.49
40 - 49	76240	.81	1311.74	.86	.92	.93	1525.91	.27	.44
50 - 59	52318	.82	1342.76	.82	.90	.93	1567.78	.22	.39
>= 60	32981	.83	1427.60	.74	.85	.94	1663.07	.17	.34
country (native speakers)									
Australia	21561	.77	1270.03	.60	.74	.93	1419.00	.14	.27
Canada	35846	.78	1289.47	.71	.82	.92	1460.42	.17	.32
United Kingdom	95536	.78	1269.58	.86	.92	.93	1422.86	.28	.46
USA	269759	.78	1302.85	.95	.97	.92	1487.41	.53	.71
education (native speakers)									
Primary	4516	.70	1316.33	.23	.36	.86	1479.12	.04	.13
Secondary	109823	.75	1332.15	.86	.93	.90	1501.97	.28	.46
Bachelor	203662	.78	1287.15	.93	.96	.92	1452.60	.46	.65
Master	89546	.80	1273.38	.87	.92	.93	1449.42	.29	.46
PhD	34359	.82	1225.81	.72	.83	.94	1414.98	.18	.32
gender (native speakers)									
Female	249704	.78	1282.91	.94	.97	.92	1456.20	.51	.69
Male	198274	.78	1299.76	.92	.96	.92	1471.47	.43	.62
handedness (native speakers)									
Right-handed	393589	.78	1290.92	.96	.98	.92	1462.39	.62	.78
Left-handed	56092	.79	1279.19	.78	.87	.92	1457.45	.19	.35
level of english (all)									
I know a few words.	10337	.57	1503.94	.21	.36	.84	1525.02	.08	.18
I can have a simple conversation.	23502	.52	1652.81	.39	.54	.87	1522.92	.11	.22
I can read a simple book.	57511	.57	1606.71	.63	.76	.89	1516.33	.18	.32
I speak and read the language fluen	239257	.70	1400.31	.92	.96	.90	1497.13	.46	.65
It is my mother tongue.	415124	.79	1283.08	.97	.98	.92	1455.75	.64	.79
native language (all)									
English	455142	.78	1289.49	.97	.98	.92	1462.74	.66	.81
Dutch	16880	.70	1413.48	.50	.65	.91	1469.25	.12	.25
Finnish	12453	.71	1513.65	.47	.62	.91	1593.68	.15	.28
French	11273	.69	1366.63	.39	.54	.92	1408.80	.11	.23
German	17723	.68	1454.63	.50	.65	.92	1480.65	.13	.26
Italian	10635	.70	1448.81	.40	.55	.90	1576.51	.16	.27
Polish	22150	.56	1516.95	.46	.61	.91	1422.40	.16	.27
Spanish; Castilian	45668	.62	1470.28	.62	.75	.89	1450.70	.17	.31
Hungarian	35626	.62	1657.17	.62	.74	.88	1655.41	.21	.35
number of foreign languages (native speakers)									
0	249490	.77	1324.70	.94	.97	.92	1494.59	.50	.69
1	131154	.79	1261.56	.89	.94	.92	1430.80	.34	.52
2	48105	.80	1224.26	.77	.86	.93	1399.55	.18	.33
best foreign language (native speakers)									
Chinese	3037	.77	1118.59	.26	.40	.94	1199.71	.08	.19
French	34938	.81	1222.88	.74	.84	.94	1390.52	.17	.32
German	12247	.81	1227.76	.50	.65	.94	1418.98	.14	.26
Italian	2819	.82	1189.45	.30	.43	.94	1410.59	.15	.22
Japanese	2969	.80	1191.28	.29	.42	.94	1363.51	.09	.20
Latin	1311	.83	1129.12	.26	.37	.92	1417.72	.13	.25
Russian	1248	.81	1152.34	.26	.36	.94	1396.72	.14	.19
Spanish; Castilian	21760	.80	1245.72	.62	.75	.93	1412.93	.16	.28

The reliabilities reported in Table 1 and Table 2 show that we collected high-quality average response time estimates for a wide range of different demographic groups. In general, the reliabilities were higher for standardized reaction times than for raw reaction times. In the English test, there was a particularly large difference between words and pseudowords in terms of the reliabilities. This difference is caused by a much lower number of observations per pseudoword in this test as the total pool of pseudowords from which pseudowords were selected was much larger.

From the perspective of the analyses conducted in the current paper the most important conclusion that can be drawn from the reported reliabilities is that we obtained stable estimates of response times for a wide range of variables that should be naturally associated with increased exposure to language. For different educational levels, in both the Dutch and the English test the reliabilities were very high for all subgroups of participants with Secondary education or higher (in all cases higher than 0.85 for RTs and 0.90 for zRTs). Similarily, we obtained reliable response time estimates for different age groups in both the English and the Dutch test (in all cases above 0.74 for RTs and 0.85 for zRTs for age groups 18 – 23 and higher).

Interestingly, many non-native English speakers participated in the English test and this led to relatively reliable set of response times for native speakers of several languages. For example, the reliabilities in the case of standardized reaction times were equal or higher than 0.60 for Spanish (0.75), Hungarian (0.74), German (0.65), Dutch (0.65), Finnish (0.62) and Polish (0.61) native speakers. This allows us to conduct an analysis of the word frequency effect on non-native speakers of a large number of languages.

Correlations with existing datasets

In order to further evaluate the quality of the collected response time estimates, we also calculated their correlations with existing megastudy datasets. In particular we looked at the correlation between the standardized reaction times from the English Lexicon Project (Balota et al, 2007) and the British Lexicon Project (Keuleers et al., 2011) with the reaction times collected in the English test as well as between response times in the Dutch Lexicon Project (Keuleers et al., 2010) and the Dutch Lexicon Project 2 (Brysbaert et al., 2015) with those from our Dutch test.

We found that the correlations between the response times collected in the current study and the existing databases were high. The correlation between the subset of the current dataset for native speakers was equal to 0.73 for BLP and 0.79 for ELP in the case of the English test, and 0.70 for DLP, 0.72 for DLP2 in the case of our Dutch test. Importantly, these correlations have to be considered in the light of the internal reliability of the existing databases, which are generally in the range 0.8 – 0.9 and constitute an upper bound for the correlation which one may expect to obtain with these databases. It is also important to note that the correlation between the response times in BLP and ELP is 0.77 and between DLP and DLP2 it is 0.79. This suggests that the word knowledge task taps into very much the same word recognition processes as the lexical decision task, despite the larger stress on personal knowledge and the smaller stress on response speed.

Next, in order to investigate whether the subsets of the current data can be assumed to carry meaningful information within each

subgroup, we looked at the amount of variance explained in different subsets of the current datasets and the existing measures.

First, we took advantage of the fact that ELP and BLP data were collected in the United States and the United Kingdom respectively, and that both DLP and DLP2 data were collected in Belgium and not in the Netherlands. Although the same language is used in these pairs of countries, there is a considerable linguistic variation between them. If our dataset reflects this variability well, we would expect to find a stronger correlation between ELP and the subset of our data that was collected from the participants in the US than in the UK, but this pattern should be reversed for the BLP. This is indeed what we observed – the correlation with ELP for the participants in the US was 0.687 and 0.668 for the participants in the UK. For BLP correlations with these two subgroups were equal to 0.595 and 0.618 respectively. This analysis was based on a random sample of 21,561 sessions for the three countries from which the highest number of native speakers participated in our test (Australia, Canada, UK, US). This is a conservative approach which allows to reduce potential differences in reliabilities of different datasets associated with unequal number of sessions that we collected for each of these countries, although it lowers the correlations relative to these that could be achieved based on the full dataset.

Also for Dutch, in line with our expectations we observed that the subset of the data based on Belgian participants had stronger correlation with the DLP (r = 0.703) and DLP2 (r = 0.728) than the data collected in the Netherlands (DLP: r = 0.660; DLP2: r = 0.656). This analysis was based on the 180,248 sessions sampled for each of the countries.

Table 3. Correlations between standardized reaction times collected in the existing databases and for various subgroups in the current test. The analyses for different countries are based on a sample of 21,561 experimental sessions in each subgroup for the English test and 180,248 sessions for the Dutch test. The analyses for different age groups are based on 28,693 test for the English, 43,968 for the Dutch test.

	Correlation			*Correlation*	
English test	*BLP*	*ELP*	Dutch test	*DLP*	*DLP2*
country			*country*		
Australia	.63	.68	Belgium	.70	.73
Canada	.60	.69	Netherlands	.66	.66
United Kingdom	.62	.67			
USA	.59	.69			
age group			*age group*		
18 - 23	.66	.72	18 - 23	.72	.74
24 - 29	.64	.71	24 - 29	.69	.72
30 - 39	.64	.71	30 - 39	.67	.69
40 - 49	.63	.70	40 - 49	.64	.67
50 - 59	.61	.70	50 - 59	.62	.65
>= 60	.59	.68	>= 60	.58	.62

Finally, because all the existing datasets are based on experiments in which the vast majority of participants were young adults, we wanted to look whether age-related differences can also be found in our dataset. For this purpose, we sampled 43,968 sessions for each age group in the Dutch test and 28,693 sessions for each group in the English test. As shown in Table 3., we indeed found a stronger correlation between the data collected from young participants in our test compared to older participants.

SIZE OF THE FREQUENCY EFFECT IN DIFFERENT FREQUENCY BANDS

After establishing the quality of the collected datasets we conducted the critical analyses of how the word frequency effect varies with exposure. We assumed that the amount of linguistic exposure should vary with the three demographic variables that we collected. First of all, we assumed that the group with the least exposure to language are non-native speakers of this language. This assumption is confirmed by relatively low accuracy in the tests collected from these participants (see Table 1). Because we collected a sufficiently large dataset from non-native speakers only for English, we conducted this particular analysis only on the English dataset.

Frequency measures

For all analyses of the English data we used an extended version of the SUBTLEX-US (Brysbaert & New, 2009) corpus including 385 million words. The corpus was created by downloading 204,408 documents from the Open Subtitles website (http://opensubtitles.org) whose language was tagged as English by the contributors of that website. Next, we removed all subtitle-related formatting from the files. To eliminate all documents that contained a large proportion of text in a language other than English, we calculated preliminary word frequencies based on all documents, and removed all documents if the 30 most frequent words did not cover at least 30% of the total number of tokens in that subtitle file. Because many subtitles are available in multiple versions we implemented

duometer[4], a tool for detecting near-duplicate text documents using the MinHash algorithm (Broder, 1997). The final version of the corpus contained 69,382 documents and 385 million tokens. We will refer to this corpus as SUBTLEX-US-V2.

The Dutch corpus was created by downloading 52,209 subtitle files from the same source. We applied the same cleaning procedure as in the case of the English subtitle corpus. The final Dutch subtitle corpus contained about 26,618 documents and 130 million tokens. As the created corpus is an extended version of the SUBTLEX-NL (Keuleers et al., 2010), we will refer to that corpus as SUBTLEX-NL-V2.

Native vs non-native speakers of a language

Because only in the English test a significant number of participants specified that they are non-native speakers of this language, this analysis was not done on the Dutch test.

In order to look at the frequency effect we conducted a piecewise regression with raw reaction times (RTs) and standardized reaction times (zRTs) as well as accuracies as dependent variables, akin to the one that we used in the corpus simulations. We split the frequency range in the same way as in that case, low frequency range included all words with Zipf frequency below 2, medium frequency from 2 to 4 and high frequency above 4 (for an illustration see Figure 6).

[4] We released *duometer* as an open-source project. The tool and its source code are available at: http://github.com/pmandera/duometer/

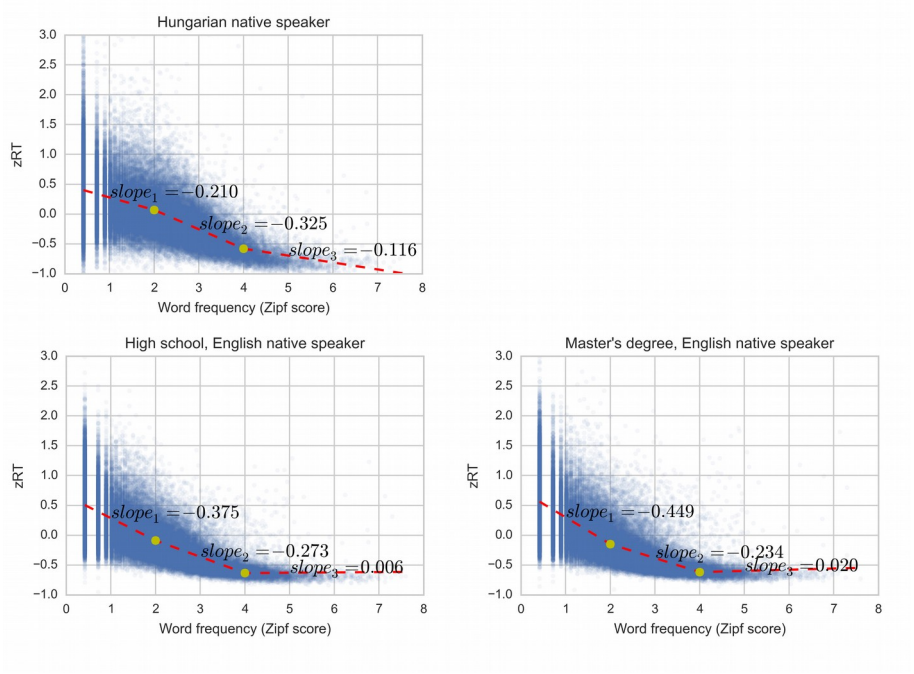

Figure 6. Word frequency effect in three groups of participants with different levels of exposure: non-native speakers of English (top panel), native speakers with secondary education (bottom, left panel) and native speaker with Master's degree (bottom, right panel). The slope in the low frequency range becomes more negative in more proficient groups but the opposite tendency is observed in the medium and high frequency ranges.

Because in our dataset the number of sessions collected for different languages differed significantly, which could bias the size of the effects for groups for which we had more observations as data for these groups can be expected to be more reliable, we considered the coefficients for differing numbers of sessions. For each group we started by randomly selecting just 1000 sessions and then increasing the number of sessions used to calculate averaged RTs, zRTs and accuracies by 1000 until a total number of sessions for this group was reached.

The coefficients in different frequency ranges for native and non-native speakers of different languages are shown in Figure 7.

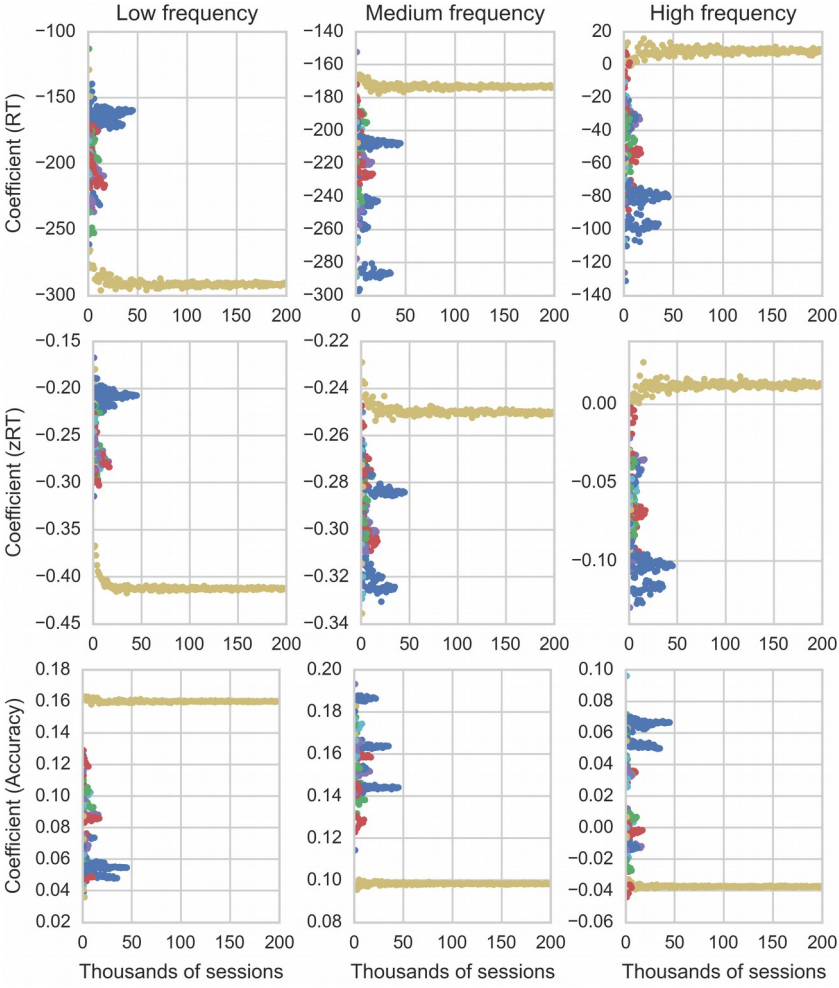

Figure 7. Slopes of the word frequency effect in different frequency ranges, as a function of cumulative number of sessions used, for native speakers of English (yellow) and other languages (other colors; due to a large number of languages the detailed legend is not shown). In the high frequency range the slope is close to 0 for native speakers but a considerable effect can be observed for the non-native speakers. The effect is larger also for the non-native speakers of English. The reversed pattern is observe in the high frequency range.

As predicted by a power function, in the highest frequency range we found a much stronger effect for non-native speakers. In this group the average coefficient calculated for the maximum number of sessions for each non-native language was equal to -49.80 for RTs, -0.06 for zRTs and 0.01 for accuracies. For the native speakers the typical word frequency effect was not observed in that frequency range, in fact a small reversed relationship was observed (8.22 in the case of RTs, 0.01 for zRTs and -0.04 for accuracies). In the medium frequency range the effect was generally much stronger than in the high frequency range. The slopes were again steeper for non-native speakers (-222.29 on average for RTs, -0.29 for zRTs and 0.15 for accuracies) compared to the native speakers (-173.26 for RTs, -0.25 for zRTs and 0.09 for accuracies). In the lowest frequency range however we observed a much stronger effect for the native speakers of the language; for native speakers in the case of all the sessions the coefficient was equal to -291.76 for RTs, -0.41 for zRTs and 0.16 for accuracies, while for the non-native speakers it averaged -198.90 in the case of RTs, -0.25 for zRTs and 0.08 for accuracies for the total number of sessions for each individual language.

Effects of education

Next, we conducted a similar analysis based on subsets of the data corresponding to different education groups. In this analysis only data collected from English native speakers from the English test and the Dutch native speakers in the Dutch test were included. For the English data groups with Secondary, Bachelor, Master and PhD educational levels were included in the analyses. As there were fewer than 5000 sessions collected from participants with Primary education in the English test we did not include it in the analyses. For Dutch, the

option to specify PhD as an educational level was not available in the questionnaire administered before the test.

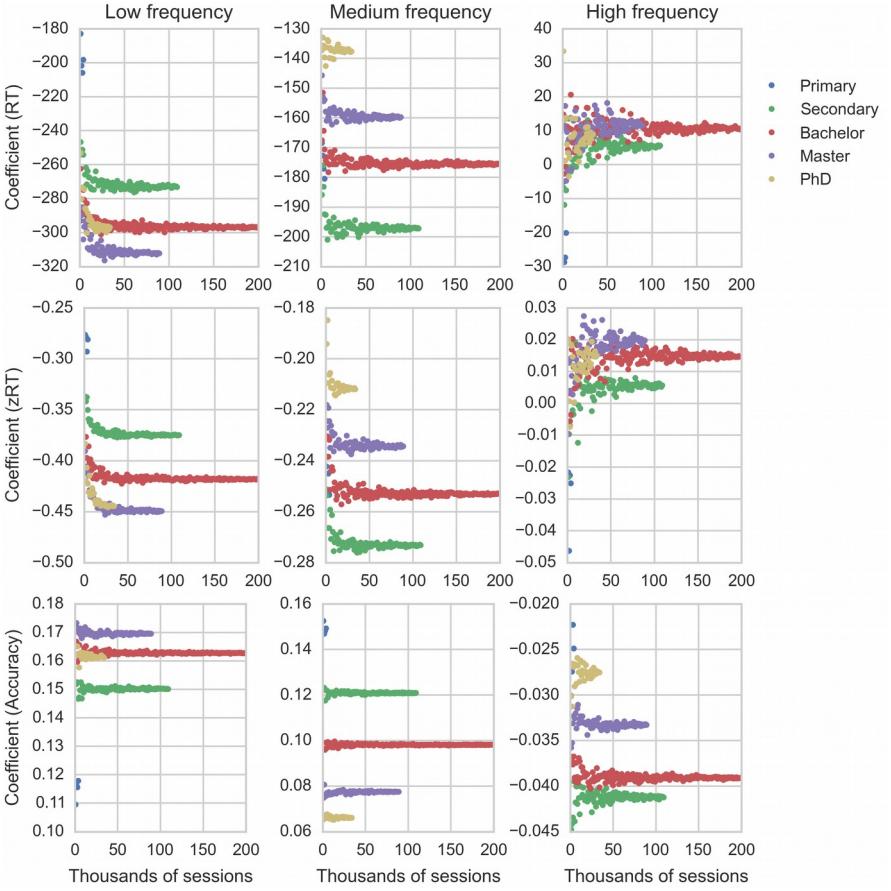

Figure 8. Slopes of the word frequency effect in different frequency ranges, as a function of cumulative number of sessions used, for different educational levels of English native speakers (English test).

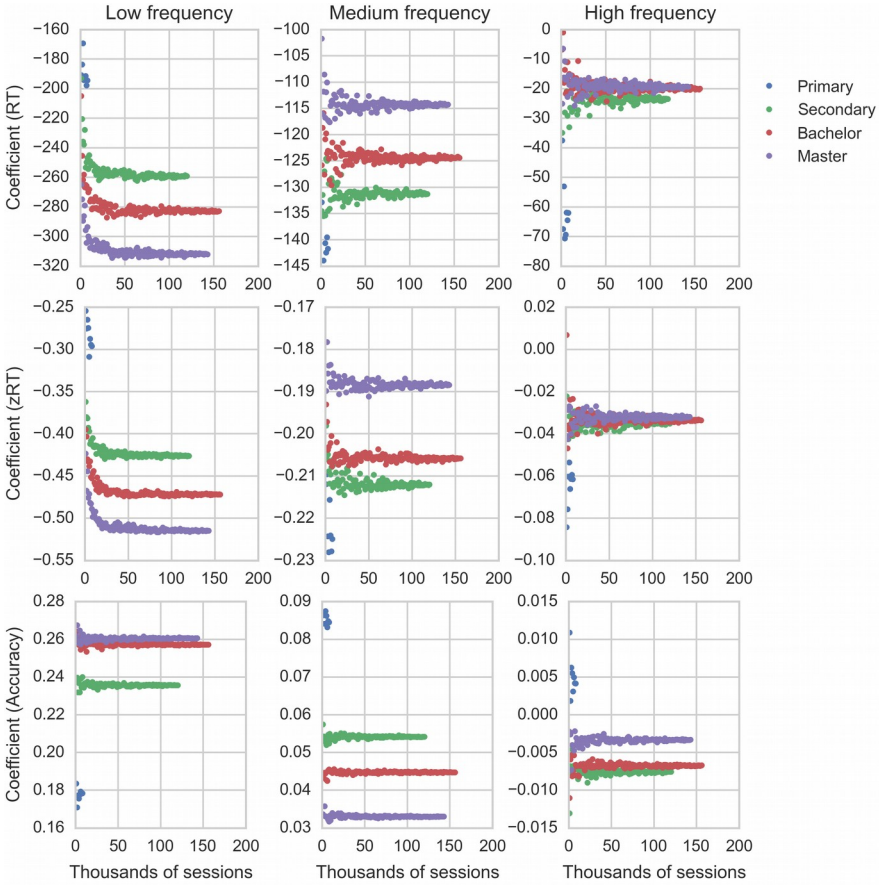

Figure 9. Slopes of the word frequency effect in different frequency ranges, as a function of cumulative number of sessions used, for different educational levels of Dutch native speakers (Dutch test).

As could be expected based on the analysis reported above, the slopes in the highest frequency range for all education groups in English test were very close to 0 or revealed a weak reversed pattern of word frequency. In the Dutch test weak effects of frequency were observed in this frequency range for response times. The slopes were steepest for participants with only Primary education (-62.00 for RTs, -0.06 for zRTs) followed by those collected from participants with Secondary education (-23.48 for RTs, -0.04 for zRTs), Bachelor's

degree (-20.01 for RTs, -0.04 for zRTs) and Master's degree (-19.37 for RTs, -0.04 for zRTs).

In the medium frequency range the slopes systematically decreased with increasing educational level. In the English test the slopes were the least steep for participants that specified PhD as their educational level (-137.63 for RTs, -0.21 for the zRTs, and 0.07 for accuracies), followed by the slopes in the dataset from participants holding a Masters's degree (-159.62 for RTs, -0.23 for zRTs and 0.08 for accuracies), Bachelor's degree (-175.59 for RTs, -0.25 for zRT and 0.10 for accuracies) and then for participants with secondary education (-197.25 for RTs, -0.27 for zRTs and 0.12 for accuracies).

We did not collect enough data from participants with only primary education to observe a meaningful pattern in that case

In this frequency range, in the Dutch test the least steep slopes were observed for participants holding Master's degree (-114.25 for RTs, -0.19 for the zRTs and 0.03 for accuracies), then Bachelors degree (-124.41 for RTs, -0.21 for zRTs and 0.04 for accuracies), secondary education (-131.28 for RTs, -0.21 for zRTs and 0.05 for accuracies) and finally primary education (-141.75 for RTs, -0.22 for z RTs and 0.08 for accuracies).

Similarly, to the analysis based on data from native and non-native speakers, the pattern of the size of the effect generally reversed in the lowest frequency range. In the case of the English test, a regular pattern of an increasing frequency effect was observed for participants with a Master's degree (-312.29 for RTs, -0.45 for zRTs, 0.17 for accuracy), Bachelor's degree (-296.84 for RTs, -0.42 for zRTs and 0.16 for accuracy), and secondary school education (-273.12 for RTs, -0.37 for zRTs and 0.15 for accuracies). The only exception to this trend was

observed for participants with a PhD degree in case who showed a slightly weaker effect compared to participants holding a Master's degree(-296.95 for RTs, -0.45 for zRTs and 0.16 for accuracies).

In the Dutch test the pattern was completely regular also in the lowest frequency range, the strongest effects were observed for participants with Master's degree (-312.08 for RTs, -0.51 for zRTs, and 0.26 for accuracies), followed by those holding Bachelor's degreee (-282.94 for RTs, -0.47 for zRTs and 0.26 for accuracies), secondary education (-259.18 for RTs, -0.43 for zRTs and 0.24 for accuracies), and primary education (-194.51 for RTs, -0.30 for zRTs and 0.18 for accuracies).

Age related differences

Finally, we looked at whether changes in the amount of exposure associated with age are also reflected in the size of the word frequency effect. Again, only data collected from native speakers were used in this analysis.

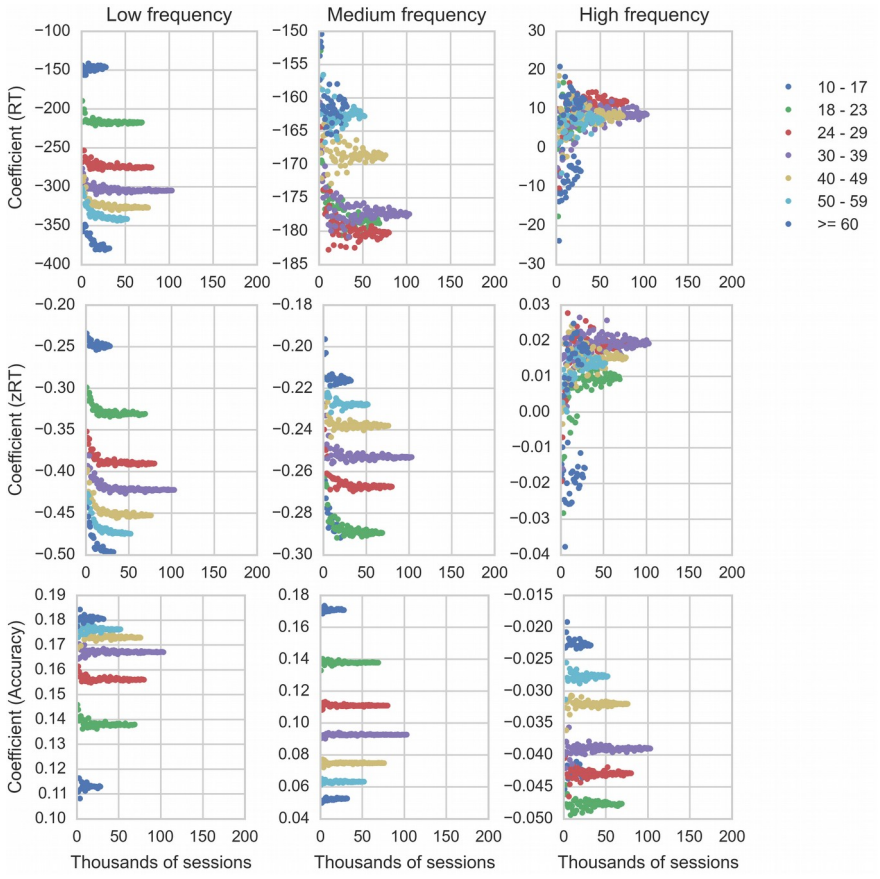

Figure 10. Slopes of the word frequency effect in different frequency ranges, as a function of cumulative number of sessions used, for different age groups of English native speakers (English test).

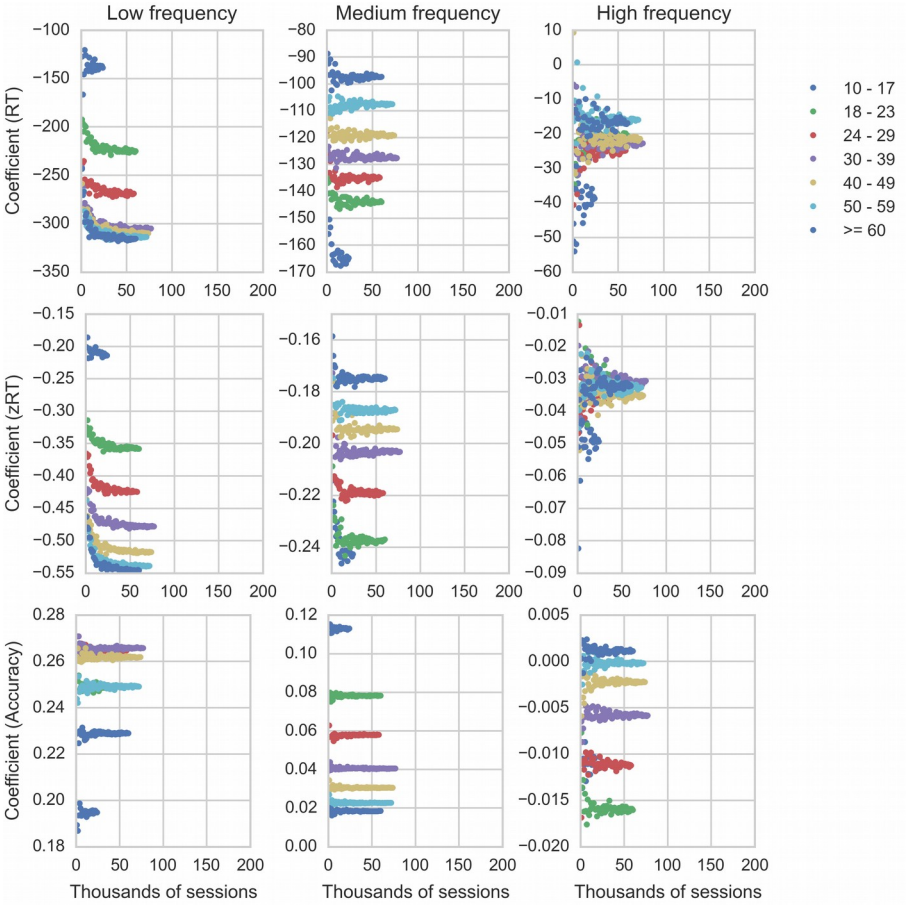

Figure 11. Slopes of the word frequency effect in different frequency ranges, as a function of cumulative number of sessions used, for different age groups of Dutch native speakers (Dutch test).

The slopes for the maximum number of sessions in each age group is reported in Table 4.

Table 4. Slopes of word frequency effect in various age groups.

		RT			zRT			Accuracy		
	N sessions	LF	MF	HF	LF	MF	HF	LF	MF	HF
English										
10 - 17	28000	-146.39	-163.07	-5.98	-0.25	-0.29	-0.02	0.11	0.17	-0.04
18 - 23	69000	-217.25	-178.85	8.46	-0.33	-0.29	0.01	0.14	0.14	-0.05
24 - 29	80000	-275.14	-180.24	11.90	-0.39	-0.27	0.02	0.16	0.11	-0.04
30 - 39	103000	-305.09	-177.42	8.58	-0.42	-0.25	0.02	0.17	0.09	-0.04
40 - 49	76000	-326.88	-168.62	8.11	-0.45	-0.24	0.02	0.17	0.07	-0.03
50 - 59	52000	-341.57	-162.76	7.20	-0.47	-0.23	0.01	0.18	0.06	-0.03
>= 60	32000	-379.36	-161.30	12.43	-0.50	-0.22	0.02	0.18	0.05	-0.02
Dutch										
10 - 17	24000	-138.78	-164.76	-38.67	-0.21	-0.24	-0.05	0.19	0.11	-0.01
18 - 23	60000	-225.06	-143.96	-20.49	-0.36	-0.24	-0.03	0.25	0.08	-0.02
24 - 29	58000	-268.91	-134.90	-24.84	-0.42	-0.22	-0.03	0.26	0.06	-0.01
30 - 39	77000	-305.18	-127.56	-22.78	-0.48	-0.20	-0.03	0.27	0.04	-0.01
40 - 49	74000	-309.92	-119.19	-21.48	-0.52	-0.19	-0.04	0.26	0.03	0.00
>= 60	60000	-315.73	-97.40	-16.89	-0.55	-0.17	-0.03	0.23	0.02	0.00

Again, as could be expected, the frequency effect was not present in the highest frequency range for most age groups, for both English and Dutch. One more time we observed a regular pattern of slopes associated with word frequency becoming less steep in the medium frequency range and increasing slopes in the lowest frequency range. This pattern was completely regular for both RTs and zRTs in the case of the Dutch test. In the English test, the pattern was as expected in the lowest frequency range for both measures of response times. In the medium frequency range, the two youngest groups had less steep slopes than would be expected for raw RTs but this irregularity disappeared completely in the case of standardized RTs.

Additional analyses

The two irregularities in the otherwise regular pattern were found in the English test in RTs associated with age (for young age groups) in the medium frequency range and for both RTs and zRTs in the case of education (PhD holders) in the low frequency range. If the hypothesis regarding the changes of the word frequency effect is correct, the observed patterns would be explained due to the fact that

we chose the boundaries between frequency ranges arbitrarily. It is possible that for the youngest participants, with relatively little exposure, the same process that is increasing the effect in the high frequency range may also play a role in the medium frequency range. For the PhD holders on the other hand, who we can assume to have the most exposure in the investigated groups, the point at which the effect already starts to decrease may have been reached also in the low frequency range.

If this hypothesis is true we should be able to adjust the irregularities by shifting the boundary between the low and the medium frequency range to higher values for the age related effects and towards lower values for the education related effects.

In the case of the age effect, after shifting the boundary to a Zipf value of 2.5, we observed a completely regular pattern of decreasing steepness of the word frequency effect with increasing age group from -165.19 for the age group 17 – 21 to -109.46 for the age group >= 60. The pattern was also perfectly regular in the lowest frequency range where the steepness increased with age from -148.89 for the age group 17 – 21 to -343.51 for the age group >= 60.

For the analysis of the educational levels, we first shifted the boundary to a Zipf value of 1.5, which resulted in a weaker irregularity but the slope in the low frequency range was still larger for the participants with Master's education (-351.71) compared to those with PhD (-339.15), in the medium frequency range the effect was as expected. After shifting the boundary to Zipf value 1, the difference in the low frequency range became even smaller, (-440.73 for PhD , -446.43 for Master education). If we looked at the same analysis conducted on standardized reaction times, the pattern was

completely regular with the slope becoming steeper with increased education from -0.47 for Secondary education to -0.64 for Master's education and -0.65 for PhD. The pattern was also completely regular in the medium frequency range, with the steepness decreasing with education from -0.30 for Secondary education to -0.28 for PhD education.

CORPUS SIZE AND VARIANCE EXPLAINED IN DIFFERENT GROUPS

A recent explanation regarding the observed interaction between skill and proficiency, proposed by Kuperman and Van Dyke (2014), attributed the interaction to the fact that larger corpora, which are often used to calculate word frequencies, overestimate the relative frequencies of low-frequency words in smaller samples, which can be assumed to be representative of the language exposure of less proficient readers. The straightforward prediction of this hypothesis is that smaller corpora should be better at predicting variance for less proficient groups of participants. Because this hypothesis does not make distinct predictions regarding separate frequency ranges, we conducted an analysis in which we fitted linear regression models without splitting the sets of words depending on their frequency range. We used a standard linear regression with standardized response times derived from different age groups in English and Dutch with a linear and quadratic effect of word frequency. Critically, we varied the size of the corpus that was used to calculate word frequencies. For English, we used subsets of SUBTLEX-US-V2 including 5, 10, 30, 50, 80, 100, 150, 200, 250, 300, 350 and 384 million words and for Dutch 5, 10, 30, 50, 80, 100 and 130 million word subset of SUBTLEX-NL-V2. Next, we looked at the amount of variance explained in each

group by the frequencies derived from different corpus sizes. To equate the number of experimental sessions used in each age group for the English test we sampled 28 thousand sessions. In the Dutch test we sampled 24 thousand sessions for each of the levels. For the analysis of educational levels we sampled 34 thousand sessions in the English test and 120 thousand sessions in the Dutch test.

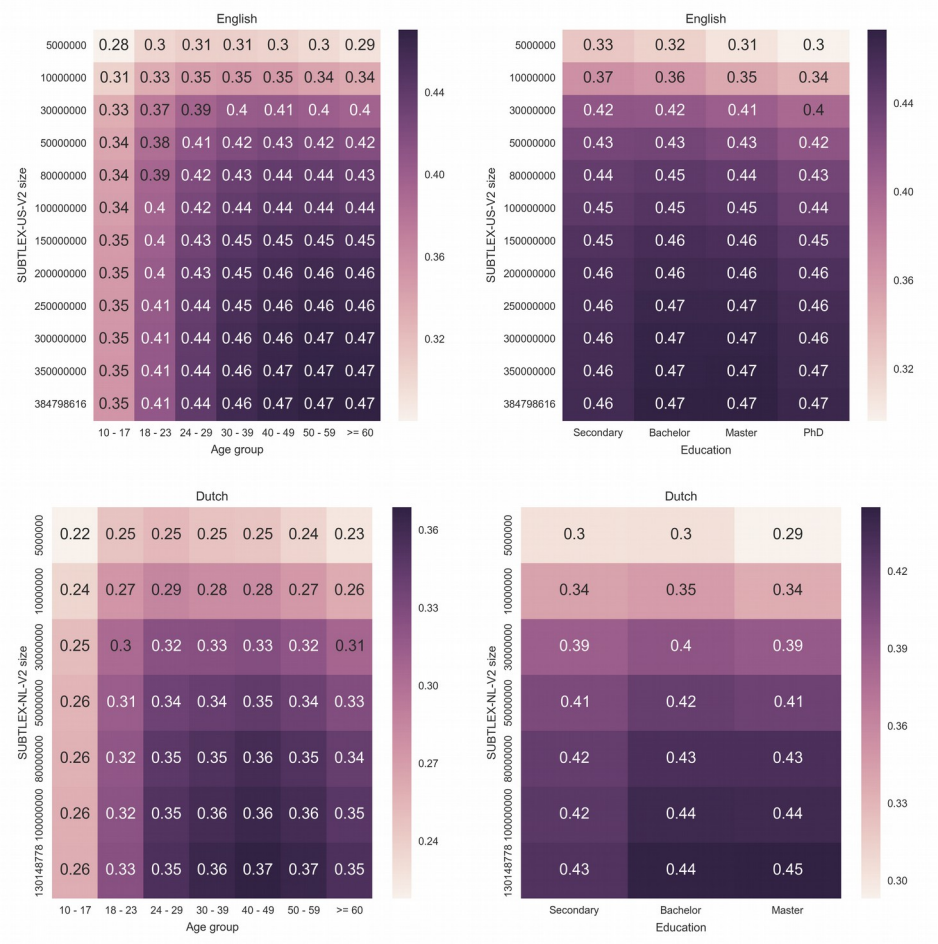

Figure 12. Percentage of explained variance in standardized reaction times in different age (left panels) and educational levels (right panels) in the English (top panels) and Dutch (bottom panels) test depending on corpus size used to calculate word frequencies.

The results of this analysis are shown in Figure 12.

In contrast to what would be expected if overestimation would be the only reason for the frequency x skill interaction, the amount of variance explained never decreased with increasing corpus size. On the other hand, based on the predictions presented in the current paper we expect the amount of explained variance to increase with increasing corpus size. Critically, because the word frequency effect decreases in the high frequency range with increased exposure, while the frequency effect increases in the low frequency range, it becomes increasingly important to have precise word frequency estimates in the low frequency range for more proficient participants and this requires a larger corpus. Hence, we expect that the size of the corpus is more important for the amount of variance explained in older and more highly educated participants. This is indeed the pattern that we observed in the data. For younger participants the percentage of explained variance became saturated very fast with increasing corpus size, however this was not the case for older participants. In fact, based on the observed trends we could expect that for the groups with the most exposure, the amount of explained variance could further increase if the size of the corpus was further increased beyond what was available to us.

DISCUSSION

In this paper, we have shown that the shape of the word frequency curve and its changes associated with proficiency can be explained purely as a function of exposure if we assume that the time to respond is described by a power function. First, we derived theoretical predictions regarding the changes in the word frequency

effect associated with increased exposure. Next, we confirmed these predictions using a corpus simulation. Finally, the predicted patterns were confirmed in behavioral data collected in two massive word knowledge experiments for Dutch and English with almost 1.5 million participants. The predicted pattern was confirmed based on a large spectrum of educational levels and, also, for the first time, when an evolution of the word frequency effect over the entire lifespan was analyzed.

The revealed pattern supports the practice effect as a primary source of the frequency effect. As predicted by Murray and Forster (2014), for such a function the word frequency effect decreases with increasing amount of exposure. However, if one jointly considers two properties of the Zipfian distribution – extreme differences in how often we experience different words and how little experience we have with the low frequency words in general – it becomes clear that the effect does not have to completely disappear but may even be increasing in the low frequency range.

In a sense, the pattern of increasing word frequency effect in the low frequency range could be interpreted as an artifact, because we have fewer observations and less data in that frequency range. However, this phenomenon is also a natural consequence of how words are distributed in a language and could be observed even if there are no other factors affecting word recognition than the sample size as shown by our corpus simulation.

Interestingly, the power function as a basic principle describing the practice effect, allows us to explain a range of phenomena observed in behavioral data. Firstly, it accounts for the shape of the frequency curve and in particular for the facts that the steepness of the

frequency effect increases with deceasing frequency range, and that the frequency effect in the low frequency range can be observed only in the lowest-proficiency speakers. Secondly, it explains why a stronger word frequency effect has been observed for non-native speakers of a language. Thirdly it explains the proficiency x skill interaction in general (and why the frequency effect is strongest in the low frequency range).

Yet another interesting property of the power function is that it becomes linear in log-log scale. In other words, researchers that are log-transforming reaction times and then using log-transformed frequencies as a predictor, are already implicitly applying a power function to their data.

The power function of practice also offers an alternative explanation of the proficiency x skill interaction to the one based on attributing it to an overestimation of the frequency estimates in smaller language samples (less proficient participants, Kuperman & Van Dyke, 2013). In this case we would expect to find smaller corpora to predict performance of less proficient participants better than larger corpora but this was not observed in the analyzed dataset: the larger corpora were always better than smaller corpora for all proficiency groups. Yet another pattern emerged from our analyses: due to a combination of the overall decrease of the effect in the high frequency range and an increase of the effect in the low frequency with increased exposure, the observable word frequency effect shifts to the lower frequency range. Because the estimation of the frequencies of the low frequency range is especially affected by the size of the corpus, the size is especially important for more proficient age groups. Nevertheless, we do not deny that using subjective word frequencies

may correct for this effect, but it has to be kept in mind that the subjective ratings of frequency may also be prone to the same power function practice effect as behavioral measures.

The described shift in the word frequency effect also means that, methodologically speaking, it may be difficult to talk about low- and high-frequency words without referring to the details of the discussed participants. Moreover, it seems that larger corpora may be necessary to model the word frequency effect in more proficient participants.

An important methodological innovation of this paper was the web based collection of data. The massive amount of data collected shows how efficient this method can be. The collected dataset seems to be of a very high quality in terms of its reliability and patterns of correlation with existing databases. Most importantly, it covers a very wide range of demographic groups and a very large part of the lexicon. However, it has to be kept in mind that comparisons of absolute reaction times between different groups should be made with caution as the technical details which may have affected the absolute values of response time measurements may systematically vary across demographic groups.

Another particularity of the current dataset is that it is based on a word knowledge task. This fact was reflected in longer overall reaction times than in typically used, speeded lexical decision tasks. However, the high correlations with existing sets of lexical decision data provide evidence that the word knowledge task has a large similarity to the lexical decision task, so that the observed pattern of results most probably generalizes to that task.

A natural question to ask in the context of the current findings is which of the existing models of word recognition could account for the power function transformation of word frequencies. A rigorous derivation of the predictions of the functional shape of the frequency effect for several models was conducted by Adelman and Brown (2008). It showed that at least some existing models, such as the instance model and its variants (Logan, 1988), predict such a functional relationship between word frequencies and response times, while other models, such as the Bayesian reader model (Norris, 2006) and the DRC (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) would not predict this kind of relationship. Further work should look at whether learning-based models of word acquisition, can account for the patterns shown in the current paper. For example, it is known that the Rescorla-Wagner model can account for an asymptotic shape of the learning curve (Miller, Barnet, & Grahame, 1995). However, it would be interesting to see if modern incarnations of this model, aimed particularly at language processing (see Baayen, Milin, Filipovic Durdevic, Hendrix, & Marelli, 2011) and considered in the context of cognitive aging (Ramscar et al., 2014), would predict the observed patterns.

It is also important to further investigate the details of the processes involved for the very low frequency words. Diependaele et al. (2012) have shown that the responses are becoming very noisy in that frequency range. Nevertheless, we observed very strong effects in the low frequency range for groups with sufficiently high proficiency.

Finally, we do not claim that the shape of the practice curve, as it is observed for word frequencies is exactly described by a power function. In fact, it is well known that averaging a mixture of

exponentials can give rise to a power function (Newell & Rosenbloom, 1981; Newman, 2005). It has also been observed that the exponential function is more adequate for describing practice effects when data from individual participants are considered (Heathcote, Brown, Mewhort, 2000). We do not want to make claims about what is the best function to describe the practice effect at the level of individual speakers and it is possible that the shape that we observed resembles a power function only because we considered averaged responses. It also has to be kept in mind that there is a wide spectrum of functions hat are intermediate between exponential and power function (Newell & Rosenbloom, 1981). What is critical here is that the practice effect cannot be excluded as a primary source of the word frequency effect in general and of the changes in the effect in groups of participants with different language exposure in particular. Nevertheless, it would certainly be interesting to re-examine the issue of an exact functional shape of the practice effect using currently available, large datasets of behavioral data.

ACKNOWLEDGEMENT

This research was made possible by an Odysseus grant from the Government of Flanders.

REFERENCES

Adelman, J. S., & Brown, G. D. A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, *115*(1), 214–227. http://doi.org/10.1037/0033-295X.115.1.214

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Baayen, R. H. (2010). A real experiment is a factorial experiment. *The Mental Lexicon*, *5*(1), 149–157.

Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*, 438-482.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General*, *133*(2), 283–316. http://doi.org/10.1037/0096-3445.133.2.283

Balota, D. A., & Ferraro, F. R. (1996). Lexical, sublexical, and implicit memory processes in healthy young and healthy older adults and in individuals with dementia of the Alzheimer type. *Neuropsychology*, *10*(1), 82.

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: what do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual Word Recognition Volume 1: Models and Methods* (pp. 90–115). Hove, East Sussex: Psychology Press.

Blevins, J., Milin, P., & Ramscar, M. (n.d.). Zipfian discrimination. *NetWordS 2015 Word Knowledge and Word Usage*, 29.

Broder, A. Z. (1997). On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings* (pp. 21–29). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=666900

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3(3), 296– 322.

Brown, S., & Heathcote, A. (2003). Averaging learning curves across and within participants. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 11–21.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. http://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(3), 441–458. http://doi.org/10.1037/xhp0000159

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. http://doi.org/10.3758/s13428-013-0403-5

Chateau, D., & Jared, D. (2000). Exposure to print and word recognition processes. *Memory & Cognition*, *28*(1), 143–153. http://doi.org/10.3758/BF03211582

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*, 661–703. http://doi.org/10.1137/070710111

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204 –256.

Cop, U., Keuleers, E., Drieghe, D., & Duyck, W. (2015). Frequency effects in monolingual and bilingual natural reading. *Psychonomic Bulletin & Review*. http://doi.org/10.3758/s13423-015-0819-2

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3), e57410. http://doi.org/10.1371/journal.pone.0057410

Diependaele, K., Brysbaert, M., & Neri, P. (2012). How Noisy is Lexical Decision? *Frontiers in Psychology*, *3*. http://doi.org/10.3389/fpsyg.2012.00348

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *The Quarterly Journal of Experimental Psychology*, *66*(5), 843–863. http://doi.org/10.1080/17470218.2012.720994

Duyck, W., Vanderelst, D., Desmet, T., & Hartsuiker, R. J. (2008). The frequency effect in second-language visual word recognition. *Psychonomic Bulletin & Review*, *15*(4), 850–855. http://doi.org/10.3758/PBR.15.4.850

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google* (pp. 47–54). Retrieved from http://www.researchgate.net/profile/Adam_Kilgarriff/publication/237127354_Proceedings_of_the_4_th_Web_as_Corpus_Workshop_(WAC4)_Can_we_beat_Google/links/00b7d5290647fbc33f000000.pdf#page=53

Garlock, V. M., Walley, A. C., & Metsala, J. L. (2001). Age-of-Acquisition, Word Frequency, and Neighborhood Density Effects on Spoken Word Recognition by Children and Adults☆. *Journal of Memory and Language*, *45*(3), 468–492. http://doi.org/10.1006/jmla.2000.2784

Gerlach, M., & Altmann, E. G. (2013). Stochastic Model for the Vocabulary Growth in Natural Languages. *Physical Review X*, *3*(2). http://doi.org/10.1103/PhysRevX.3.021006

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008b). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis☆. *Journal of Memory and Language*, *58*(3), 787–814. http://doi.org/10.1016/j.jml.2007.07.001

Gollan, T. H., Slattery, T. J., Goldenberg, D., Van Assche, E., Duyck, W., & Rayner, K. (2011). Frequency drives lexical access in reading but not in speaking: The frequency-lag hypothesis. *Journal of Experimental Psychology: General*, *140*(2), 186–209. http://doi.org/10.1037/a0022256

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207.

Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. The Hague: Mouton.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology, 41*, 401-410.

Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, *52*(12), 5186–5201. http://doi.org/10.1016/j.csda.2007.11.008

Keuleers, E., & Balota, D. A. (2015). Megastudies, Crowdsourcing, and Large Datasets in Psycholinguistics: An Overview Of Recent Developments. *The*

Quarterly Journal of Experimental Psychology, *68*(8), 1457–68. http://doi.org/10.1080/17470218.2015.1051065

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. http://doi.org/10.3758/BRM.42.3.627

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. http://doi.org/10.3758/BRM.42.3.643

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice Effects in Large-Scale Visual Word Recognition Studies: A Lexical Decision Study on 14,000 Dutch Mono- and Disyllabic Words and Nonwords. *Frontiers in Psychology*, *1*. http://doi.org/10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2011). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. http://doi.org/10.3758/s13428-011-0118-4

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1665–92. http://doi.org/10.1080/17470218.2015.1022560

Kolers, P. A. (1975). Memorial consequences of automatized encoding. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(6), 689–701. http://doi.org/10.1037/0278-7393.1.6.689

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. http://doi.org/10.3758/s13428-012-0210-4

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(3), 802–823. http://doi.org/10.1037/a0030859

Laubrock, J., Kliegl, R., & Engbert, R. (2006a). SWIFT explorations of age differences in eye movements during reading. *Neuroscience & Biobehavioral Reviews*, *30*(6), 872–884. http://doi.org/10.1016/j.neubiorev.2006.06.013

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(1), 12–31. http://doi.org/10.1037/0278-7393.34.1.12

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492.

McCusker, L. M. (1977, April). Some determinants of word recognition: Frequency. Paper presented at the 24th Annual Convention of the Southwestern Psychological Association, Fort Worth, TX.

Miller, R.R., Barnet, R.C., Grahame, N. J. (1995). Assessment of the Rescorla-Wagner Model. *Psychological Bulletin*, 117 (3), 363-386.

Murray, W. S., & Forster, K. I. (2004). Serial Mechanisms in Lexical Access: The Rank Hypothesis. *Psychological Review*, *111*(3), 721–756. http://doi.org/10.1037/0033-295X.111.3.721

Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition*, *28*(5), 832–840.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review, 113*, 327– 357.

O'Dowd, S. (1984). Does vocabulary decline qualitatively in the old age? *Educational Gerontology*, 10, 357–368.

Pugh, K. R., Frost, S. J., Sandak, R., Landi, N., Rueckl, J. G., Constable, R. T., … Mencl, W. E. (2008). Effects of stimulus difficulty and repetition on printed word identification: An fMRI comparison of nonimpaired and reading-disabled adolescent cohorts. *Journal of Cognitive Neuroscience*, *20*(7), 1146–1160.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The Myth of Cognitive Decline: Non-Linear Dynamics of Lifelong Learning. *Topics in Cognitive Science*, *6*(1), 5–42. http://doi.org/10.1111/tops.12078

Rayner, K., Reichle, E. D., Stroud, M. J., Williams, C. C., & Pollatsek, A. (2006). The effect of word frequency, word predictability, and font difficulty on the eye movements of young and older readers. *Psychology and Aging*, *21*(3), 448–465. http://doi.org/10.1037/0882-7974.21.3.448

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, *47*(2), 309–327. http://doi.org/10.3758/s13428-014-0471-1

Shaywitz, S., Shaywitz, B., Fulbright, R., Skudlarski, P., Mencl, W., Constable, R., …, Gore, J. C. (2003). Neural systems for compensation and persistence: Young adult outcome of childhood reading disability. *Journal of Biological Psychiatry, 54*, 25–33.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3(3), 271–295.

Spieler, D. H., & Balota, D. A. (2000). Factors influencing word naming in younger and older adults. *Psychology and Aging*, *15*(2), 225–231. http://doi.org/10.1037//0882-7974.15.2.225

Spiess, A.-N., & Neumeyer, N. (2010). An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC Pharmacology*, *10*(1), 6.

Tainturier, M.-J., Tremblay, M., Lecours, A. (1989). Aging and the word frequency effect: A lexical decision investigation. *Neuropsychologia, 27 (9),* 1197-1202.

Van Wijnendaele, I., & Brysbaert, M. (2002). Visual word recognition in bilinguals: Phonological priming from the second to the first language. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(3), 616–627. http://doi.org/10.1037//0096-1523.28.3.616

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–207. http://doi.org/10.3758/s13428-012-0314-x

Whitford, V., & Titone, D. (2012). Second-language experience modulates first- and second-language word frequency effects: Evidence from eye movement

measures of natural paragraph reading. *Psychonomic Bulletin & Review*, *19*(1), 73–80. http://doi.org/10.3758/s13423-011-0179-5

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53–79. http://doi.org/10.1037/a0024177

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Cambridge: Addison-Wesley.

Chapter 5. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation[1]

ABSTRACT

Recent developments in distributional semantics (Mikolov et al., 2013) include a new class of prediction-based models that are trained on a text corpus and that measure semantic similarity between words. We discuss the relevance of these models for psycholinguistic theories and compare them to more traditional distributional semantic models. We compare the models' performances on a large dataset of semantic priming (Hutchison et al., 2013) and on a number of other tasks involving semantic processing and conclude that the prediction-based models usually offer a better fit to behavioral data. Theoretically, we argue that these models bridge the gap between traditional approaches to distributional semantics and psychologically plausible learning principles. As an aid to researchers, we release semantic vectors for English and Dutch for a range of models together with a convenient interface that can be used to extract a great number of semantic similarity measures.

[1] This chapter is based on a paper accepted for publication in Journal of Memory and Language pending minor changes as Mandera, P., Keuleers, E., & Brysbaert, M. (accepted). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*.

INTRODUCTION

Distributional semantics is based on the idea that that words with similar meanings are used in similar contexts (Harris, 1954). In this line of thinking, semantic relatedness can be measured by looking at the similarity between word co-occurrence patterns in text corpora. In psychology, this idea inspired a fruitful line of research starting with Lund and Burgess (1996) and Landauer and Dumais (1997). The goal of the present paper is to incorporate a new family of models recently introduced in computational linguistics and natural language processing research by Mikolov and colleagues (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Mikolov, Chen, Corrado, Dean, 2013) into psycholinguistics. In order to do so, we will discuss the theoretical foundation of these models and evaluate their performance on predicting behavioral data on psychologically relevant tasks.

COUNT AND PREDICT MODELS

Although there are different approaches to distributional semantics, what they have in common is that they start from a text corpus and that they often represent words as numerical vectors in a multidimensional space. The relatedness between a pair of words is quantified by measuring the similarity between the vectors representing these words.

The original computational models of semantic information (arising from the psychological literature) were based on the idea that the number of co-occurrences of words in particular contexts formed the basis of the multidimensional space and that the vectors were obtained by applying a set of transformations to the count matrix. For

instance, Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) starts by counting how many times a word is observed within a document or a paragraph. The Hyperspace Analogue to Language (HAL; Lund & Burgess, 1996) counted how many times words co-occurred in a relatively narrow sliding window, usually consisting of up to ten surrounding words. Because of the common counting step, following Baroni, Dinu & Kruszewski (2014) we will refer to this family of models as ***count models***.

In count models, the result of this first step is a word by context matrix. What usually follows is a series of transformations applied to the matrix. The transformations involve some kind of a weighting scheme, based on frequency-inverse document frequency, positive pointwise mutual information (PPMI), log-entropy, and/or a dimensionality reduction step (most commonly singular value decomposition; SVD). Sometimes the transformation is the defining component of the method, as is the case for LSA, which is based on SVD. In other cases, however, the transformations have been applied rather arbitrarily to the counts matrix based on empirical studies investigating which transformations optimized the performance on a set of tasks. For example, in its original formulation, the HAL model did not involve complex weighting schemes or dimensionality reduction steps, but later it was found that they improved the performance of the model (e.g., Bullinaria & Levy, 2007, 2012). Transformations are now often applied when training models (e.g., Recchia & Louwerse, 2015; Mandera, Keuleers, & Brysbaert, 2015).

If we consider Marr's (1982) distinction between computational, algorithmic, and implementational levels of explanation, the count models are *only defined* at the *computational*

level (Landauer & Dumais, p. 216): They consist of functions that map from a text corpus to a count matrix and from the count matrix to its transformed versions. Regarding the algorithmic level, Landauer and Dumais (1997) did not attribute any realism to the mechanisms performing the mapping. They only proposed that the counting step and its associated weighting scheme could be seen as a rough approximation of conditioning or associative processes and that the dimensionality reduction step could be considered an approximation of a data reduction process performed by the brain. In other words, it cannot be assumed that the brain stores a perfect representation of word-context pairs or runs complex matrix decomposition algorithms in the same way that digital computers do.[2] In the case of HAL, even less was said about the psychological plausibility of the selected algorithms. Another problem is that count models require all the information to be present before the transformations are applied, whereas, in reality, learning in cognitive systems is incremental, not conditional on the simultaneous availability of all information.

[2]It is known that dimension reduction can be performed by biological (e.g. Olshousen & Field, 1996) and artificial (Hinton & Salakhutdinov, 2006) neural networks. This fact is rarely mentioned when authors discuss various approaches to distributional semantics in the psycholinguistic literature.

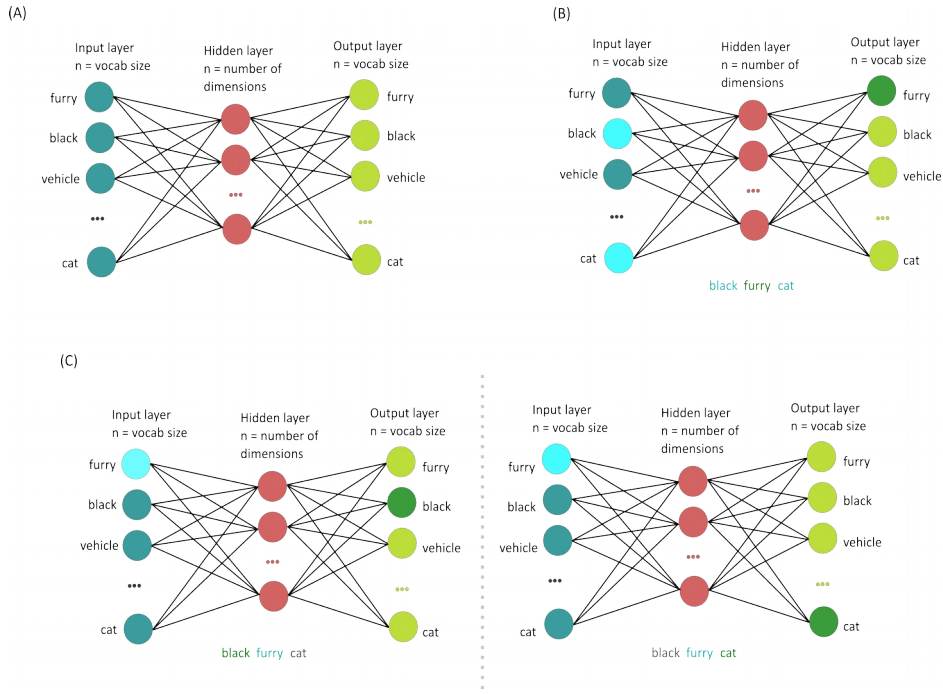

Figure 1. Both the CBOW and the skip-gram models are simple neural networks (a) composed of the input, the hidden and the output layer. In the input and the output layers each node corresponds to a word so the number of nodes in these layers is equal to the total number of entries in the lexicon of the model. The number of nodes in the hidden layer is a parameter of the model. The training is performed by sliding a window through a corpus and adjusting the weights to better fit training examples. When the model encounters a window including a phrase black furry cat, the CBOW model (b) represents the middle word furry by an activation of the corresponding node in the output layer and all context words (black and cat) are simultaneously activated in the input layer. Next, the weights are adjusted based on the prediction error. In the case of the skip-gram model (c) the association between each of the context words (black and cat) is predicted by the target word (furry) in a separate learning step. When training is finished, the weights between the nodes and the input layer and the hidden nodes are exported as the resulting word vectors

In other words, although the count models, like all computational models, were very specific about which properties were extracted from the corpus to build the count matrix, and which mathematical functions were applied to the counts matrix in the transformation step, they made it much less clear how these computations could be performed by the human cognitive system.[3] This is surprising, given that the models originated in the psychological literature.

Unexpectedly, a recent family of models, which originated in computer science and natural language processing, may be more psychologically plausible than the count models. Mikolov and colleagues (2013a) argued that a relatively simple model based on a neural network (see Figure 1) can be surprisingly efficient at creating semantic spaces.

This family of models is built on the concept of prediction. Instead of explicitly representing the words and their context in a matrix, the model is based on a relatively narrow window (similar in size to the one often used in the HAL model) sliding through the corpus. By changing the weights of the network, the model learns to predict the current word given the context words (Continuous Bag of Words model; CBOW) or the context words given the current word (skip-gram model). Because of the predictive component in this family of models, again following Baroni et al. (2014), we will refer to these models as ***predict models***. As indicated above, there are two main types: the CBOW model and the skip-gram model.

[3]Although Landauer and Dumais (1997) discuss how the LSA algorithm could hypothetically be implemented in a neural network, this aspect is not reflected in their implementation of the model.

Even though the predict models originated outside the context of psychological research and were not concerned with psychological plausibility, the simple underlying principle – implicitly learning how to predict one event (a word in text corpus) from associated events–, is arguably much better grounded psychologically than constructing a count matrix and applying arbitrary transformations to it. The implicit learning principle is congruent with other biologically inspired models of associative learning (Rescorla & Wagner, 1972), given that they both learn on the basis of the deviation between the observed event and the predicted event (see Baayen, Milin, Filipovic Durdevic, Hendrix, & Marelli, 2011). An additional advantage of the model is that it is trained using a stochastic gradient descent, which in this case means that it can be trained incrementally with only one target-context pairing to be available for each update of the weights, and does not require all co-occurrence information to be present simultaneously as is the case with the count models.

To illustrate in what sense we consider the predict models to be psychologically plausible, we would like to compare them to the Rescorla-Wagner model – a classical learning model (for a review see Miller, Barnet, & Grahame, 1995), which has also been successfully applied to psycholinguistics (Baayen et al., 2011). This model learns to associate cues with outcomes by being sequentially presented with training cases. For each training case, if there is a discrepancy between the outcomes predicted based on current association weights and the observed outcomes (lack of an expected outcome or presence of an unexpected outcome), the weights are updated using a simple learning rule.

Interestingly, the update rule of the Rescorla-Wagner model is known to be mathematically equivalent to the delta rule (Sutton and Barto, 1981), which describes stochastic gradient descent in a neural network composed of a single layer of connections and which was independently proposed outside of the context of psychological research (Widrow-Hoff, 1960). The same rule has been generalized to networks consisting of multiple layers of connections and non-linear activation functions as a backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986) and is used to determine changes in connection weights in connectionist models. In other words, the Rescorla-Wagner model is just a special case of the backpropagation algorithm used with a stochastic gradient descent.

Similarly to the Rescorla-Wagner model, the learning mechanism which is used to train the predict models is also based on backpropagation with stochastic gradient descent. These models learn to minimize errors between the outcomes predicted based on the cues and the observed outcomes by updating the weights of the connections between the nodes in the network when observing events in a text corpus. Here cues and outcomes correspond to target and context words in a sliding window, and each update of the weights is based on a predicted and observed pairing between the target word and its context. The learned semantic representation, which can be thought of as a pattern of activation of the hidden nodes for a word in an input layer, is learned as a by-product of learning to associate contexts and target words. The model is usually trained in one pass over the corpus with the number of the training cases being dependent on the size of the corpus.

In this sense, the predict models are trained using a similar technique as the Rescorla-Wagner learning rule, adapted for a network which includes a hidden layer and a non-linear activation function. It could be argued that introducing the hidden layer and non-linearity to the model make it conceptually more complex than the Rescorla-Wagner model.[4] However, it is clear that it may be impossible to represent more complex phenomena, such as semantics, in models as simple as the Rescorla-Wagner model. In the case of the predict models, the hidden layer is necessary to introduce a dimensionality reduction step (Olshousen & Field, 1996; Hinton & Salakhutdinoy, 2006) and a non-linear (softmax) activation function is necessary to transform activations of outcomes to probabilities. In fact it has been argued that using neural networks deeper than three layers may be necessary and justified to simulate and explain cognitive phenomena: deep neural networks have proven to be successful in a large variety of fields (for a review see Le Cun, Bengio, & Hinton, 2015) and hierarchal processing is also recognized as a fundamental principle of information processing in the human brain (Hinton, 2007). The need for recognizing deeper architectures as valid approaches to cognitive modeling has also been proposed in the psychological literature (Zorzi, Testolin, & Stoianov, 2013; Testolin, Stoianov, Sperduti, Zorzi, 2015).

[4] It is important to note that although a network with no hidden layers may be simpler conceptually, it does not necessarily mean that it is more parsimonious in terms of the number of parameters that need to be specified. For example, consider a network with 50,000 words as cues and the same number of outcomes. A fully-connected network with a single layer of connections, such as the Rescorla-Wagner model, would require 50,000 x 50,000 = 2.5 billion parameters (weights) to be specified, while introducing a hidden layer including 300 nodes drastically reduces this number to 2 x 50,000 x 300 = 30 million parameters (weights).

In addition to their potential theoretical appeal, the predict models were shown to offer a particularly good performance across a set of tasks and generally outperform the count models (Baroni et al., 2014; Mandera et al., 2015) or perform as well as the best tuned count model. On the other hand, it has also been argued that the superior performance of the predict models is largely due to using better tuned parameters as default for training these models than is the case for the count models (Levy, Goldberg, Dagan, & Ramat-Gan, 2015). Even if the performance of the predict models does not surpass that of the count models, they are generally much more compact in terms of how much computational resources they require, which is also of practical importance.

Although the predict models are built on a quite simple principle, it is not as obviously clear as in the case of the count models what, in mathematical terms, these models are computing (Goldberg & Levy, 2014). Interestingly, it has been argued that some of the predict models may implicitly perform a computation that is mathematically equivalent to the dimensionality reduction of a certain type of the count model. In particular, Levy and Goldberg (2014) argued that the skip-gram model is implicitly factorizing a PMI transformed count matrix shifted by a constant value. If this is the case, and the relationship between the two classes of models becomes well understood, this could create an interesting opportunity for psychologists by showing how mathematically well-defined operations (PPMI, SVD) can be realized on psychologically plausible systems (neural networks) to acquire semantic information.

Given the potential convergence of the predict and count models it becomes especially important to introduce the predict

models to psycholinguistics. If the count models are well specified at Marr's (1982) computational level of explanation, the predict models could provide an algorithmic level explanation, bringing us closer to understanding how semantic representations may emerge from incrementally updating the predictions about co-occurrences of events in the environment. Nevertheless, because to our knowledge this is the first time these models are discussed in a psycholinguistic context, in the current paper we did not focus on investigating the convergence of the two classes of models but chose to train different semantic spaces with typical parameter settings and details of the training procedures.

To advance our understanding of the new predict models (both CBOW and skip gram) and their relationship to the more traditional count models in a psychological context, we performed an evaluation of the three types of models against a set of psychologically relevant tasks. In order to gain a more complete picture of how these models perform we tried to explore their parameter space instead of limiting ourselves to a single set of parameters. In addition, we wanted to find out how much of what we have learned about count models can be generalized to the predict models.

Of course, the investigated implementations of the predict models are only loosely related to psychologically plausible principles (such as prediction). We do not claim that the investigated predict models represent a human capacity to learn semantics in a fully realistic way, but rather we argue that they should be investigated carefully because they may represent an interesting starting point for bridging the theoretical gap between the count models, various transformations applied as part of these models, and fundamental psychological principles.

COMPARING DISTRIBUTIONAL MODELS OF SEMANTICS

There is a rich literature in which different approaches to distributional semantics have been evaluated. In general they form two types of investigations: Either various parameters and transformations within one approach are tested to find the most successful set of parameter settings (e.g. Bullinaria and Levy 2007, 2012), or different approaches are compared to each other to establish the best one (e.g. Baroni et al., 2014; Levy et al., 2015).

The evaluations are often based on a wide range of tasks. For example, Bullinaria and Levy (2007, 2012) compared the performance of a HAL-type count model on four tasks: The Test of English as a Foreign Language (TOEFL; Landauer & Dumais, 1997), distance comparison, semantic categorization (Patel et al., 1997; Battig & Montague, 1969), and syntactic categorization (Levy et al., 1998). The authors varied a number of factors such as the window size, the applied weighting scheme, whether dimensionality reduction was performed, whether or not the corpus was lemmatized (all inflected words replaced by their base forms), and so on. They found that the best results on their battery test were achieved by the models that used narrow windows, the PPMI weighting scheme, and a custom, SVD-based dimensionality reduction step. The lemmatization or use of stop-words did not improve the performance of the model.

Comparisons of different classes of models include a recent comparison of the predict approach to the traditional count model on a range of computational linguistic benchmark tasks: Baroni et al. (2014) compared the models using semantic relatedness (Rubenstein and Goodenough, 1965; Agirre et al., 2009, Bruni et al., 2014), synonym detection (TOEFL; similar to Landauer and Dumais, 1997),

concept categorization (purity of clustering categorization, Almuhareb, 2006; Baroni et al., 2008; Baroni et al., 2010), selection preferences (noun-verb pairs, how similar are they as subject-verb or object-verb pairs, Baroni and Lenci, 2010; Padó & Lapata, 2007; McRae et al., 1998), and analogy (Mikolov et al., 2013a) and found that the predict models had a superior performance on computational linguistic benchmark tasks and were more robust to varying parameter settings. Levy, Goldberg, Dagan, & Ramat-Gan (2015) show that although count models lack the robustness of predict models, they can work equally well with specific weighting schemes and dimensionality reduction procedures.

It is clear that the benchmark tasks from computational linguistics may not be the most relevant ones for issues related to human semantic processing and representation. For instance, a lot of attention has been devoted to how well various distributional semantic models perform on the TOEFL, which consists of choosing which of four response alternatives most closely matches a target word over 80 trials with increasing difficulty. Unless we want to model scholastic over-achievement, there is no a priori reason to believe that the model scoring best on this test is also the psychologically most plausible one. A simple psycholinguistic benchmark could consist of correctly predicting the proportion of alternatives chosen by participants. In this respect, the relatedness ratings or elicited associations tasks used in the computational linguistics benchmarks can also be considered valid benchmarks for psycholinguistics. However, evaluating computational models in psycholinguistics also involves comparing predictions about the time course associated with processing stimuli. The most frequently used task to study the time course of semantic processing in

humans is semantic priming. This task consists of the presentation of a prime word followed by a target stimulus. Usually, the task involves either reading the target word out loud (naming) or deciding whether the stimulus is an existing word or a pseudoword (lexical decision). The task does not involve an explicit response about the semantic relationship between prime and target. However, it is assumed that the time it takes to read the word out loud or to make a decision on its lexicality is decreased by the degree of semantic relatedness between the prime and the target. Therefore, in contrast to other benchmarks in which participants are asked to give explicit responses about semantic content, semantic priming is assumed to inform us about the implicit working of semantic memory.

PREDICTING SEMANTIC PRIMING WITH DISTRIBUTIONAL MODELS

The question of whether semantic similarity measures derived from distributional semantics models can predict semantic priming in human subjects has been investigated in a number of psycholinguistic studies. In terms of the employed methodology these investigations can be divided in two classes. Some studies simply look at the stimuli across related and unrelated priming conditions and investigate whether there is a significant difference in semantic space derived similarity scores between these conditions. Other studies try to model the semantic priming at the item level by means of regression analysis.

The first class of studies is exemplified by Lund, Burgess, & Atchley (1995) who found that the HAL-derived similarity measures significantly differed for semantically related and unrelated conditions. A similar approach was taken by McDonald & Brew

(2004) and Pado & Lapata (2007), who used distributional semantics models to model semantic priming data from Hodgson (1991). Jones, Kintch, & Mewhort (2006) compared the BEAGLE, HAL and LSA models on a wide range of priming tasks, and investigated differences in how well these methods mimicked the results of multiple priming studies.

The regression-based approach was already employed in Lund and Burgess (1996), who reported that relatedness measures derived from HAL significantly correlated with semantic priming data from an existing priming study (Chiarello, Burgess, Richards, & Pollock, 1990). A detailed examination of the factors modulating the size of the semantic priming effect based on 300 pairs of words was conducted by Hutchison, Balota, Cortese, & Watson (2008). In a regression design the authors found no effect of the LSA score. However, it is worth noting that a large number of other predictors were entered in the analysis, including other semantic variables, such as forward and backward association strength from an association study by Nelson, McEvoy, & Schreiber (1998). Collinearity of these measures may have contributed to the fact that no significant effect of the LSA score was found. In addition, the null result does not prove that computational indices are unable to predict semantic priming, as the quality of the used semantic space may have been suboptimal.

Another item-level study was conducted recently by Günther, Dudschig and Kaup (2016) in German. In that study the authors carefully selected a set of items spanning the full range of LSA similarity scores computed on the basis of a relatively small corpus of blogs (about 5 million words). The authors found a small but significant effect of the LSA similarity scores. The critical difference

between this study and the one conducted by Hutchison et al. (2008) was in how the authors analyzed the data: Hutchison et al. (2008) first subtracted RTs in the related condition from the RTs in the unrelated contition and then fitted regressions to the resulting difference. Günther et al. (2015) simply predicted the reaction times to the target words while including a set of other variables (including semantic similarity with the prime) as predictors. Difference scores between correlated variables are known to have a low reliability (Cronbach & Furby, 1970) and arguably reduced reliability may have contributed to lack of significant effect in the study by Hutchison et al. (2008).

Although the item-level, regression based approach has multiple advantages over factorial designs (Balota, Cortese, Sergent-Marshall, Spieler, & Yap 2004; Balota, Yap, Hutchison, Cortese, 2012), until recently it was difficult to conduct this type of analysis on a sufficiently large number of items. Fortunately, due to the recent rise of megastudies (Keuleers & Balota, 2015), the situation is improving rapidly. Thanks to the semantic priming project (SPP) ran by Hutchison and colleagues (2013), we now have a much better opportunity to look at how much of the total variability in primed lexical decision times (LDT) and word naming times can be explained by semantic variables based on distributional semantics models. The advantage of this approach is that with enough data we can directly model RTs as a function of semantic similarity between the prime and the target, also including other critical predictors known to influence performance on psycholinguistic tasks. Because in a megastudy approach it is natural to focus on effect sizes more than on categorical decisions based on statistical significance, the method lends itself to

comparing various semantic spaces by examining how much variance in RTs they account for.

Corpus Effects In Distributional Semantics

The performance of distributional semantics models in accounting for human data can be affected by the degree to which the training corpus of the model corresponds to the input human participants have been exposed to. Ideally, the model would be trained on exactly the same quality and size of data as participants of psycholinguistic experiments (typically first-year university students). Of course, this ideal can only be approximated. In particular, much of the language humans have been exposed to is spoken and can only be used for modeling purposes after a time-consuming transcription process. Instead, models are typically based on written language which is available in large quantities but is often less representative of typical language input.

However, it has been observed that frequency measures based on corpora of subtitles from popular films and television series outperform frequency measures based on much larger corpora of various written sources. For instance, Brysbaert, Keuleers, and New (2011) showed that word frequency measures based on a corpus of 50 million words from subtitles predicted the lexical decision times of the English Lexicon Project (Balota et al., 2007) better than the Google frequencies based on a corpus of hundreds of billions words from books. A similar finding was reported by Brysbaert, Buchmeier, Conrad, Jacobs, Bölte, and Böhl (2011) for German. In particular, word frequencies derived from non-fiction, academic texts perform worse (Brysbaert, New, & Keuleers, 2012). On the other hand,

Mandera, Keuleers, Wodniecka, & Brysbaert (2015) showed that a well-balanced corpus of written texts from various sources performed as well as subtitle-based frequencies in a Polish lexical decision task.

An interesting question in this respect is how important the corpus size is for distributional semantics vectors. Whereas a corpus of 50 million words may be enough for frequency measures of individual words, larger corpora are likely to be needed for semantic distance measures, as estimation of semantic vectors composed of hundreds of values may be a more demanding task than assigning a frequency to a word. Some evidence along these lines was reported by Recchia and Jones (2009), who observed that using a large corpus is more important than employing a more sophisticated learning algorithm. The two corpora they compared contained 6 million words versus 417 million words. On the other hand, De Deyne, Verheyen, & Storms (2015), based on a comparison between corpus samples of various sizes, conclude that corpus size is not critical for modeling mental representations. So, in addition to the effects of size, the language register tapped into by the corpus could also influence semantic distance measures based on distributional models. We will discuss this issue by comparing the performance of models based on subtitle corpora with the performance of models based on written materials. If subtitle corpora perform better than the larger text corpora of written materials, this indicates that register is an important variable. In addition, if the concatenation of both corpora turns out to be inferior on some tasks, this is again an indication of the importance of the register captured by subtitle corpora.

Evaluating Semantic Spaces as Psycholinguistic Resources

The availability of the priming lexical decision and word naming megastudy data collected by Hutchison and colleagues (2013) makes a systematic comparison of various measures of semantic relatedness feasible and opportune. In addition to various distributional semantic models, semantic relatedness ratings can also originate from feature-based data (McRae, Cree, Seidenberg, & McNorgan, 2005), human association norms (Nelson et al., 1998), or semantic relatedness ratings (Juhasz, Lai, & Woodcock, 2015). While we will include these alternatives in our comparison, it should be noted that they have some important practical limitations: (1) they are defined only for a subset of words and (2) they do not exist in most languages that can be potentially of interest to psycholinguists.

To perform the evaluation, the logic of evaluating word frequency norms (Brysbaert and New, 2009; Keuleers et al., 2010) will be followed. In these evaluations, various word frequency norms are used to predict lexical decision and word naming RTs in order to identify the set of norms that accounts for the largest percentage of variance in the behavioral data (ideally together with other lexical variables that affect word processing times, such as word length and neighborhood density). An almost identical procedure can be applied to semantic spaces. A linear regression model can be fitted to the lexical decision and naming latencies of target words preceded by semantically related or unrelated primes. The variables known to influence word recognition (frequency, length, and similarity to other words) will be used as baseline predictors, to which the semantic distance between the prime and the target derived from the various

distributional semantics models will be added. This leads to the measurement of how much extra variance in behavioral data can be accounted for by adding relatedness measures from each distributional semantic model.

Although this approach can be informative of a model's absolute performance, it does not give an indication of the relative evidence in favor of each model. The approach based on comparing amount of variance explained is also biased towards more complex models when comparing them against the baseline (including more variables gives more explanatory power but may result in overfitting the training data). In order to overcome these limitations, we applied a regression technique based on Bayes factors (e.g. Wagenmakers, 2007) as described by Rouder and Morey (2013; see also Liang, Paulo, Molina, Clyde, & Berger, 2008). The Bayes factor is a measure of relative probability of the data under a pair of alternative models. This method also automatically incorporates a penalty for model complexity (Wagenmakers, 2007) and is flexible with respect to which models can be compared, for instance to allow for the comparison of non-nested models, which is difficult in a frequentist approach (Kass & Raftery, 1995). This property allows us to quantify the relative evidence in favor of different models including predictors derived from various semantic spaces.

Although we consider the data from the semantic priming project to be the most informative with respect to getting insight into the semantic system of typical participants in psychology experiments, we will also look at how well the various measures perform on a number of other tasks, and we will include some data from the Dutch language, to test for cross-language generalization. In addition, where

possible we will compare the outcome of the new variables to those currently used by psycholinguists.

Initially, we intended to compare two count models (LSA-inspired and HAL-inspired) with two predict models (CBOW and skip-gram). However, when we tried to calculate the LSA-type model on our corpora, it became clear that the number of documents (particularly in the UKWAC corpus) was too large to represent the term by document matrix in computer memory and perform SVD on that matrix. As a result, we had to use a non-standard, more scalable implementation of the SVD algorithm implemented in the Gensim toolkit (Rehurek & Sojka, 2010), which returned vectors that were not doing particularly well. Because it is not clear whether the bad performance of the LSA-type measure is due to the inferior performance of the LSA approach itself or to the algorithm, and because LSA-based measures in the past have done worse than HAL-based measures, we decided not to include the former in the analyses reported below. For the most important task (semantic priming), however, we do provide the LSA measures as provided by the Colorado website for comparison purposes.

Finally, to obtain a more nuanced view of how the models perform across different parameter settings we explore their parameter space. By doing so, we make sure that we give each model maximal opportunity and we can examine whether all models are similarly affected by, for instance, the size of the window around the target word or the number of dimensions included in the model.

CURRENT STUDY

For each corpus the tokenization was done by extracting all the alphabetical strings. Following Bullinaria and Levy (2007, 2012) no lemmatization or exclusion of function words was used. To represent the degree to which two words are related according to the used semantic spaces we computed cosine distances between word vectors *u* and *v* according to the formula:

$$D_{\cos}(u,v)=[1-\frac{u\cdot v}{||u||||v||}]$$

In this formula $u \cdot v$ stands for a dot product between vectors *u* and *v*, and $||u||$ and $||v||$ for the length of the vector *u* and *v* respectively.

ENGLISH

Text corpora

The corpora we used for creating the English semantic spaces were UKWAC (a corpus of about 2 billion words resulting from a web crawling program; Ferraresi, Zanchetta, Baroni, & Bernardini, 2008) and a corpus of about 385 million words compiled from film and television subtitles. More information about UKWAC can be found in Ferraresi et al. (2008).

The subtitle corpus was created based on 204,408 documents downloaded from the Open Subtitles website (http://opensubtitles.org) whose language was tagged as English by the contributors of that website. We first removed all subtitle related formatting. Next, to eliminate all documents that contained a large proportion of text in a

language other than English, we calculated preliminary word frequencies based on all documents, and removed all documents in cases where the 30 most frequent words did not cover at least 30% of the total number of tokens in that subtitle file. Because many subtitles are available in multiple versions we implemented *duometer* [5], a tool for detecting near-duplicate text documents using the MinHash algorithm (Broder, 1997). The final version of the corpus contained 69,382 documents and 385 million tokens.

We also combined the two corpora for the purpose of computing the semantic spaces. The combined corpus contained 2.33 billion tokens and 2.76 million documents.

Model training

We trained the (HAL-type) count model by sliding a symmetrical window through the corpus and counting how many times each pair of words co-occurred. We considered the 300,000 most frequent terms in the corpus as both target and context elements (Baroni et al., 2014). Next, we transformed the resulting word by word co-occurrence matrix using the positive pointwise mutual information (PPMI) scheme (Bullinaria & Levy, 2007). The transformation involved computing pointwise mutual information (Church & Hanks, 1990) for each pair of words x and y according to the formula:

$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

[5]We released *duometer* as an open-source project. The tool and its source code are available at: http://github.com/pmandera/duometer

Where *p(x)* is the probability of the word *x* in the text corpus, *p(y)* is the probability of the word *y* in the text corpus and *p(x, y)* is the probability of the co-occurrence of the words *x* and *y*. In the final step, the values of the cells in the matrix for which the pointwise mutual information values were negative were substituted with 0, so that the matrix contained only non-negative values (hence *positive* pointwise mutual information).

We trained the CBOW and skip-gram models using Gensim (Rehurek & Sojka, 2010)[6], an implementation that is compatible with word2vec (Mikolov et al., 2013a) – the original implementation of the predict models. For these models, all word forms occurring minimally 5 times in the corpus were included. Each model was trained using 50, 100, 200, 300 and 500 dimensions. We set the parameter k for negative sampling to 10 and the sub-sampling parameter to 1e-5. Sub-sampling is a method of mitigating the influence of the most frequent words (Mikolov et al., 2013a) by randomly removing words with a probability higher than a pre-specified threshold. Negative sampling is a computational optimization that avoids computing probabilities for all words in an output layer. In each learning case only a subset of words is considered.

An important parameter influencing the performance of count models (Bulinaria & Levy, 2007, 2012) is the size of the sliding window. We varied this parameter for the count and predict models in the range from 1 to 10 words before and after the target word[7] (i.e., the

[6] The toolkit is available at https://radimrehurek.com/gensim/

[7] The CBOW and skip-gram models limit the size of the window used on individual learning trials to a randomly chosen value in the range from 1 to the requested window size.

minimal window size of 1 included 3 words: the target word, one word before, and one word after).

Evaluation tasks

In order to keep vocabulary size constant across the count and predict models and across the three corpora used (subtitles, written texts, and their combination), we used only the subset of words that all semantic spaces had in common. We also wanted to compare our semantic spaces with the best performing space from Baroni et al. (2014; CBOW model with 400 dimensions, window size 5, negative sampling value 10, trained on the concatenation of the UKWAC, Wikipedia and the British National corpus including 2.8 billion words)[8]. Therefore, we further limited the vocabulary of the models to the intersection with the vocabulary of that dataset. The resulting semantic spaces contained 113,000 distinct words.

Semantic priming – method

We used the data from the Semantic Priming Project (Hutchison et al., 2013), which contains lexical decision times and naming times to 1,661 target words preceded by four types of primes. Two prime types were semantically related to the target but differed in their association strength; the other two types were unrelated primes matched to the related primes in terms of word length and word frequency. The Semantic Priming Project contains two more variables of interest for our purpose. They are the semantic similarity measures derived from LSA (based on the general reading ability dataset, trained on the TASA corpus, 300 dimensions; Landauer & Dumais, 1997) and from BEAGLE (Jones et al., 2006). These numbers allow

[8] Downloaded from: http://clic.cimec.unitn.it/composes/semantic-vectors.html

us to compare the newly calculated measures to the current state of the art in psycholinguistics. As the data were not available for all prime-target pairs, this further reduced the dataset. In the end 5,734 of the original 6,644 prime-target pairs remained.

For lexical decision (LDT) all non-word trials were excluded from the dataset and for both LDT and word naming we excluded all erroneous responses. We excluded all trials with RTs deviating more than 3 standard deviations from the mean and computed z-scores separately for each participant and each session. Finally, we averaged the z-scores for each prime-target pair and used the result as the dependent variable in our analyses.

Next, we fitted linear regression models with various predictors to evaluate the amount of variance in the standardized RTs that could be accounted for. First, we calculated a baseline model including log word frequency (SUBTLEX-US; Brysbaert & New, 2009), word length (number of letters), and orthographic neighborhood density (Coltheart, Davelaar, Jonasson, & Besner, 1977) of both the prime and the target (all variables as reported in the Semantic Priming Project dataset). Then, we fitted another linear regression model including the baseline predictors plus the measure of semantic distance between the prime and the target provided by the semantic space we were investigating, and looked at how much extra variance the semantic similarity estimate explained. We used all pairs of stimuli irrespective of the condition (both related and unrelated words).

Semantic priming – results

The baseline regression model including the logarithm of word frequency, length, and neighborhood density (all predictors included

for both the prime and the target word) explained 38.9% of the variance in the lexical decision RTs and 31.2% of the variance in word naming latencies (see Figure 2).

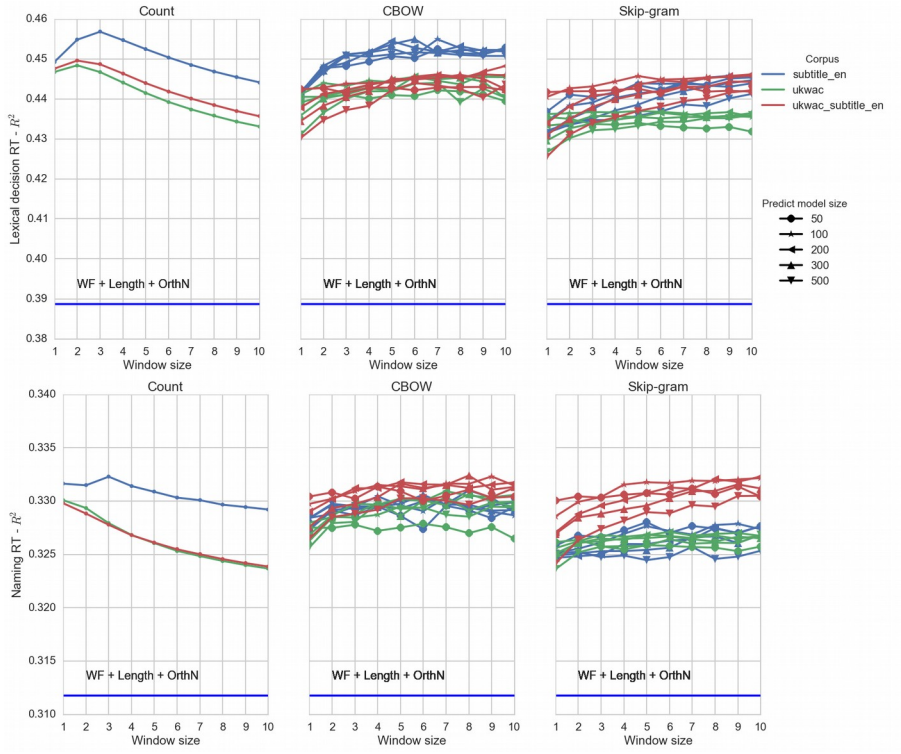

Figure 2. Performance of the three types of models on the Semantic Priming Project (Hutchison et al., 2013) dataset. The straight blue lines indicate the performance of the baseline model which did not include semantic predictors. Although the best count model in the LDT tasks performs slightly better than the best predict model (CBOW), its performance decreases rapidly with increasing window size. For naming, the predict models generally provide a better fit to behavioral data. The models trained on the subtitle corpora or on the concatenation of the subtitle corpus and the UKWAC corpus perform particularly well on these tasks.

When the relatedness scores from the distributional semantics models were added as a predictor, the amount of variance explained increased for both tasks. The improvement was already highly

significant for the relatedness measure based on the worst performing model. For LDT, this was the skip-gram model trained on the concatenation of the subtitle and the UKWAC corpus with dimensionality 500 and window size 1 [$F(1, 5729) = 367.27$, $p < 0.001$)]; for word naming it was the skip-gram model trained on the UKWAC corpus with dimensionality 500 and window size 1 [$F(1, 5729) = 99.778$, $p < 0.001$)].

On average, the models with added distance measures derived from the count model explained 44.5% (*SD* = 0.63%) of the variance in lexical decision and 32.8% (*SD* = 0.27%) in naming. The models based on CBOW similarities explained 44.5% (*SD* = 0.5%) of the variance in the primed lexical decision task and 33.0% (*SD* = 0.1%) of the variance in the naming task. The models involving the skip-gram relatedness explained 43.9% (*SD* = 0.5%) of the variance in the lexical decision reaction times and 32.7% (*SD* = 0.2%) of the variance in word naming latencies.

The best performing count model, both for lexical decision RTs and word naming latencies, was the model trained on the subtitle corpus with window size 3. It explained 45.7% of the total variance in lexical decision reaction times and 33.2% of the variance in word naming latencies.

The best performing CBOW model in lexical decision had 300 dimensions and was trained on the subtitle corpus with a sliding window of 6 words to the left and 6 words to the right. It explained 45.5% of the total variance in reaction times. For word naming, the best performing CBOW model had 300 dimensions, was trained on the concatenation of the UKWAC and the subtitle corpus using window size of 8 (33.2% of explained variance).

For the skip-gram models, the best performing model for lexical decision had 200 dimensions and was trained on the concatenation of the subtitle corpus and the UKWAC corpus using a window size of 10. It explained 44.6% of the variance. The best performing skip-gram model for naming was trained on the same corpus but had 200 dimensions and a window size of 10. It explained 33.2% of the variance in naming latencies.

Several interesting findings emerged from our analyses. First, in many cases the models trained on the subtitle corpus outperformed the models based on the UKWAC written corpus or the combination of the two corpora. This effect was particularly clear for the count models (both in LDT and word naming) and the CBOW models (in LDT). The difference was less clear for the skip-gram models. In all cases, however, the addition of 385 million words from the subtitle corpus to the 2.33 billion corpus of written texts considerably improved performance.

A second remarkable observation is that the best models are quite comparable but have different window sizes. In particular, for the count model there is a steep decrease in performance with increasing window size above 3 which was not observed for the predict models. As a result, the optimal window size is larger for the predict models than for the count model.

Semantic priming – a comparison with the existing measures of semantic similarity

To further gauge the usefulness of the new semantic similarity measures, we compared the extra variance they explained to that explained by currently used measures. The Semantic Priming Project database includes measures for LSA and BEAGLE. Currently, if a

distributional semantics model is used for the purpose of selecting experimental stimuli, psychologists tend to rely on the LSA space available through a web interface at the University of Colorado Boulder (http://lsa.colorado.edu/; Landauer & Dumais, 1997). This is understandable, as the semantic space was created to accompany a classic paper and because the resource has a practical interface which makes data extraction easy. Yet, given the recent developments in distributional semantics and the availability of much larger corpora than the one on which the CU Boulder spaces were trained (most prominently the TASA corpus of about 11 million words), there is a need to reevaluate whether the LSA-based semantic spaces should be the default choice for measuring semantic relatedness in psychological research.

The TASA-based LSA similarity scores explained 43.9% of the variance in lexical decision reaction times and 32.7% of the variance in naming. The BEAGLE scores explained 43.0% of the variance in lexical decision reaction times and 32.3% of the variance in word naming latencies.[9] All values are below those of the best performing CBOW model (45.5% in LDT and 33.2% in naming).

Our best models also compare well relative to the spaces trained by Baroni et al. (2014). The best performing semantic space of Baroni et al. (2014) explained 44.0% of the variance in lexical decision reaction times and 33.0% of the variance in word naming latencies.

To examine how much more variance could be explained by human word association norms (Nelson et al., 1998) and feature norms (McRae et al., 2005), we performed an analysis on the subsets

[9] BEAGLE scores based on cosine distances; the other measures performed worse.

of words that are included in these datasets.[10] We compared the semantic similarity indices based on the human data to those of the best count, CBOW and skip-gram spaces for the lexical decision task. There were 2,904 cue-target pairs that were simultaneously present in the priming data, the association norms and the vocabulary of our semantic spaces.

For this subset of Semantic Priming Project data, the baseline regression model including logarithm of word frequency, length and neighborhood density of both the cue and the target explained 38.9% of the variance in lexical decision times and 31.2% in word naming latencies. The model that additionally included human forward association strength explained 41.7% of the variance in lexical decision RTs and 32.7% of the total variance in word naming. The best performing count model (trained on the subtitle corpus, using window size 3) explained 42.3% of the variance in lexical decision RTs and 31.9% of the variance in word naming latencies. The best CBOW model (trained on the subtitle corpus; 300 dimensions; window size 6) accounted for 41.9% of the variance in LDT RTs and 32.0% in word naming latencies. The best skip-gram model (trained on the concatenation of the UKWAC and subtitle corpus; 200 dimensions; window 10) explained 41.0% of the variance in lexical decision and 32.1% of the variance in naming. As can be seen, all models performed very similarly and close to what can be achieved by human data. We would like to note, however, that it is harder to explain additional variance in RTs based on relatedness data, because the

[10] Similar analysis could in theory be run using the scores derived from the Simlex-999 and the Wordsim-353 ratings but the overlap with the semantic priming data was too small in these cases to allow a meaningful analysis.

subset of the Semantic Priming Project that was used for this analysis contained only pairs of words generated as associates in the Nelson et al. (1998) database, which significantly reduced the range of relatedness values.

The intersection between the feature norms from McRae et al. (2005), the semantic priming data, and the vocabulary data of our datasets included 100 word pairs. The baseline model explained 37.0% of the variance in LDT RTs and 29.3% of the variance in word naming latencies. Adding the relatedness scores computed as the cosine between the features vectors increased the percentage of variance accounted for by the model to 42.7% for LDT RTs and to 29.8% for word naming latencies. The amount of variance explained by the model in which we inserted the measures derived from the best performing count model was 54.6% for LDT RTs and 35.3% for word naming. In the case of the best CBOW model, the total explained variance amounted to 52.8% for lexical decision and 32.3% for naming. When the best performing skip-gram model word distance estimates were included in the model, it explained 52.3% of the variance in LDT RTs and 31.9% of the variance in word naming latencies. So, for this dataset, the semantic spaces actually outperformed the human data.

Semantic priming – Bayes factors analysis

For all Bayesian analyses reported in this paper we adopted an approach described by Rouder and Morey (2013; see also Liang, Paulo, Molina, Clyde, & Berger, 2008). We used default[11] mixture-of-variance priors on effect size. We also conducted a series of analyses

[11]The default 'medium' setting for the rscaleCont argument in the *regressionBF* function in the R BayesFactor package, corresponding to the *r* scale = sqrt(2)/4.

with altered priors but this did not change the qualitative pattern of results, so we report only analyses conducted with default settings.

The results of the analysis are reported in Table 1.

For both LDT and naming we first identified baseline models that included an optimal combination of lexical, non-semantic covariates. For both the prime and the target, we considered the following co-variates: log of word frequency, length, and orthographic neighborhood density. The subsequent analyses were conducted with reference to the best baseline models identified for each of the tasks. In the first analysis, to obtain the most conservative indication of whether the semantic relatedness measures reported here improve the models based on lexical predictors, we again considered all the possible submodels of the six lexical covariates with addition of the worst performing semantic relatedness measures for each task. In the case of both LDT and naming even the models including the worst performing semantic measures compared favorably to the baseline models.

Table 1. The results of the Bayes factor analysis of the English semantic priming data. Bayes factors for the baseline models are reported with reference to the intercept-only model and for the remaining models with reference to the baseline model. The worst and best relatedness measures included in the Bayesian analyses were selected separately for each task based on the R^2 in the previous analyses.

Model type	*Variables in the selected model*	*Bayes Factor*
LDT		
baseline (lexical only)	WF_{target} + len_{target} + ON_{target}	BF_{10} = 2.15 x 10^{605}
lexical + worst relatedness	WF_{target} + WF_{prime} + len_{target} + ON_{target} + rel_{worst}	$BF_{1baseline}$ = 1.24 x 10^{74}
lexical + best relatedness	WF_{target} + WF_{prime} + len_{target} + ON_{target} + rel_{best}	$BF_{1baseline}$ = 2.10 x 10^{144}
lexical + multiple relatedness	WF_{target} + WF_{prime} + len_{target} + ON_{target} + rel_{BEAGLE} + rel_{CBOW} + rel_{count}	$BF_{1baseline}$ = 4.79 x 10^{161}
Naming		
baseline (lexical only)	WF_{target} + WF_{prime} + len_{target} + len_{prime}	BF_{10} = 5.99 x 10^{457}
lexical + worst relatedness	WF_{target} + WF_{prime} + len_{target} + len_{prime} + rel_{worst}	$BF_{1baseline}$ = 1.72 x 10^{20}
lexical + best relatedness	WF_{target} + WF_{prime} + len_{target} + len_{prime} + rel_{best}	$BF_{1baseline}$ = 2.34 x 10^{36}
lexical + multiple relatedness	WF_{target} + len_{target} + len_{prime} + rel_{CBOW} + rel_{count}	$BF_{1baseline}$ = 5.50 x 10^{41}

Note. WF_{target} = $\log_{10}$ of the target word frequency; WF_{prime} = $\log_{10}$ of the prime word frequency; len_{target} = number of letters in the target word; len_{prime} = number of letters in the prime word; ON_{target} = orthographic neighborhood density of the target word; rel_{worst} = the worst relatedness measure; rel_{best} = the best relatedness measure; rel_{CBOW} = the best CBOW relatedness measure; rel_{count} = the best count measure; rel_{BEAGLE} = the relatedness measure based on BEAGLE.

Secondly, in order to establish which variables would make it to the model with the highest Bayes factor, for the two tasks we considered models including the best performing semantic relatedness measures in addition to the lexical variables. The results of this analysis can be inspected in Table 1.

In the next Bayes factor analysis, we evaluated whether different semantic spaces carry unique information that may be informative for predicting behavioral data. We simultaneously included multiple semantic relatedness measures: the Colorado LSA space, BEAGLE, the space trained by Baroni et al. as well as the best of each type of the semantic spaces (CBOW, skip-gram and count space) trained for the current study in addition to log target and prime word frequency, length of the prime and the target and orthographic neighborhood density for the target (for the sake of computational efficiency, we removed orthographic neighborhood density of the prime from the set of predictors). Interestingly, this analysis shows that the optimal model includes multiple measures of semantic similarity both in the LDT and in the naming task.

Finally, we considered the models including word association data and feature norms data as predictors in the model in addition to the lexical variables and the three best semantic relatedness measures t hat we trained. In these analyses, word association norms and feature norms were not among the most successful models, which included semantic predictors based on our semantic spaces.

Because in many cases the differences in R^2 associated with models including different relatedness measures were rather small, we directly compared models including each type of relatedness measures (count, CBOW and skip-gram). For each of the type of the models we

first considered all subsets of the lexical predictors and the relatedness measure which explained the highest percentage of the variance in the analyses above. Next, we directly compared the best models including each type of relatedness measures. In the analysis of the LDT data the best performing count model (trained on the subtitle corpus with window size 3) performed better than the best CBOW model (including 300 dimensions, trained on the subtitle corpora with window size 6; $BF_{10} = 521$) and the best skip gram model (trained on the concatenation of the UKWAC and subtitle corpora, 200 dimensions, window size 10, $BF_{10} = 5.77 \times 1024$). For naming, the best of all three types of models explained about 33.2% of the variance in lexical decision. However, when performing the Bayes factor analysis of this dataset the best model including the relatedness measures derived from the CBOW model included one more predictor (log of prime word frequency), as a result due to increased complexity it had a lower Bayes factor relative to the best model associated with the best count ($BF_{01} = 17.08$) and skip-gram models ($BF_{01} = 21.18$).

In summary, the Bayesian analysis showed overwhelming evidence in favor of including semantic relatedness measures derived from semantic spaces in both naming and lexical decision tasks, even when the worst performing of our models were evaluated. Interestingly, the optimal model included relatedness measures derived from multiple models. This suggests that different models may carry unique information that independently explains human performance in semantic priming. Finally, it seems that distributional semantics models outperform the available human associations and featural norms in explaining human performance in semantic priming.

Word association norms – method

In order to evaluate how well the different models can predict human association data we used the dataset collected by Nelson, McEvoy and Schreiber (1998). This contains word associations for 5,019 stimulus words collected from over 6,000 participants. We limited the analysis to those associations that were present in all our semantic spaces, which resulted in a dataset of 70,461 different cue-response pairs (on average 14 associates per word).

To compare the word associations generated by humans to those generated by semantic spaces, we computed a metric based on the relative entropy between the probability distribution of the top 30 associates generated by the model and the associates generated by the human participants. This metric captures not only the probabilities for the words generated by humans but also evaluates whether the same words are generated by the semantic spaces.

To calculate the metric, the following steps were followed:

1. For each semantic space, we calculated the cosine distances between the cue word and all the other words, and selected the 30 words that were nearest to the cue word. A value of 30 corresponds to about twice the number of associates that are typically generated in human data. As such, it allows for enough responses to be considered while not deviating too much from the number of associates generated by humans.
2. Next, the similarity score for each associate was normalized by dividing it by the sum of all the similarity values for the cue. The same procedure was applied to the human association data, with associate counts being converted to probabilities. If the

semantic space did not include the associate that was present in the human data or vice versa, a value of 0 was assigned.

3. Next, an additive smoothing was applied to each distribution using a smoothing term of 1/n, in which n is the number of elements in the distribution.
4. The relative entropy between probability distribution P and another probability distribution Q was computed with the formula:

$$D_{KL}(P\|Q)=\sum_i P(i)\log_2\frac{P(i)}{Q(i)}$$

5. Finally, the relative entropies were averaged across all cue words and the average relative entropy was used as the final score of a given semantic space. Note that a relative entropy measure is a measure of distance between probability distributions and, hence, the smaller the measure, the better the fit.

Word association norms - results

To compute a baseline for the performance of the models on the association norms, we used a set of semantic spaces with word vectors containing nothing but random values. The average relative entropy between the associations norms and 10 such randomly generated semantic spaces was 0.84 *(SD* = 0.0001; see Figure 3).

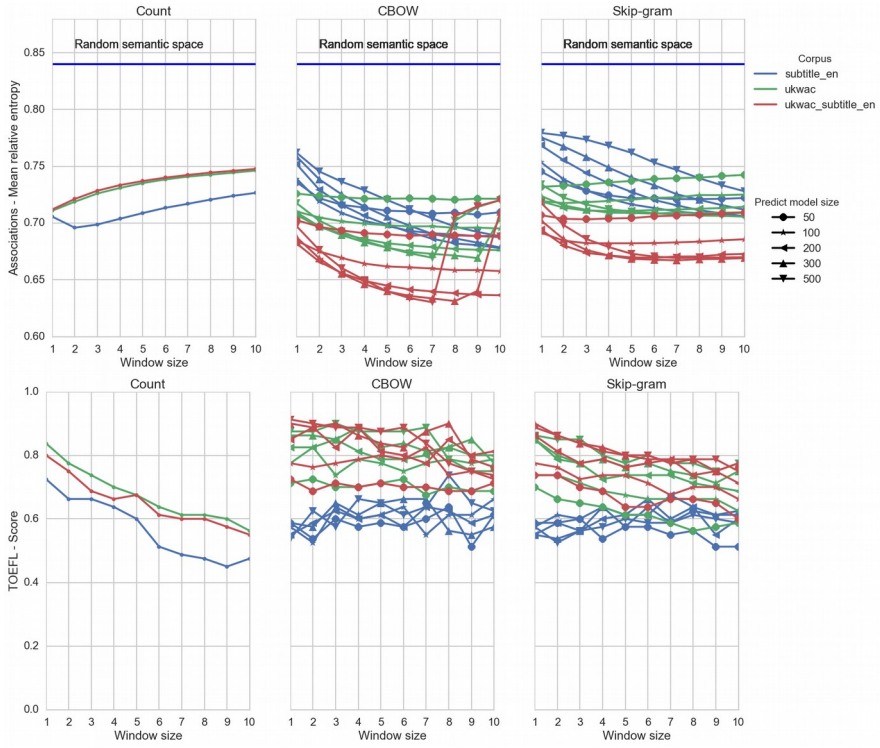

Figure 3. Performance of the three types of models on the association norms dataset (upper panels) and on TOEFL (lower panels). The predict models generally outperform the count models. Models trained on a subtitle corpus perform worse than the models trained on the UKWAC corpus or concatenation of the two corpora. Note that for the association norms lower entropy is better.

Both the predict and the count semantic spaces managed to achieve lower relative entropies than the baseline (recall that lower is better, as relative entropy measure is a measure of distance between distributions). The best performing count model was trained on the subtitle corpus using window size 2 (relative entropy = 0.70).

The best performance of the predict models was achieved by the CBOW model with 500 dimensions trained on the concatenation of the subtitle and the UKWAC corpora with window size 7, which had a relative entropy of 0.63. The best skip-gram model was trained

on the same corpus, used the same window size, but had 300 dimensions and had relative entropy of 0.66.

The average relative entropy for the measures derived from the count models was 0.73 (*SD* = 0.02). For the CBOW models it was 0.69 (*SD* = 0.03) and for the skip-gram model 0.71 (*SD* = 0.03).

Like before, the best count models were those with small window sizes, whereas small window sizes were detrimental for the predict models. On this task, CBOW performed best, followed by the skip-gram model, and finally by the HAL-type count model. For comparison, the semantic space from Baroni et al. (2014) had a relative entropy of 0.68, which was better than the average of the models evaluated here but worse than the best of those models.

Similarity/Relatedness ratings - method

We used two datasets of human judgments of semantic similarity and relatedness to evaluate semantic distance estimates on the basis of semantic spaces.

Wordsim-353 (Agirre et al., 2009) is a dataset including 353 word pairs, with about 13 to 16 human judgments for each pair. For this dataset the annotation guidelines given to the judges did not distinguish between similarity and relatedness. However, the dataset was split into a subset of related words and a subset of similar words on the basis of two further raters' judgment about the nature of the relationship for each word pair.

The second set of human judgments is Simlex-999 (Hill, Reichart, & Korhonen, 2014), which contains similarity scores for 999 word pairs. What makes it different from Wordsim-353 is its clear distinction between similarity and relatedness. In the case of Simlex-

999 participants were given very clear instruction to pay attention to the similarities between words and not to their relatedness, so that word pairs such as *car* and *bike* received high similarity scores, whereas *car* and *petrol*, despite being strongly related, received low similarity scores.

To evaluate how well each semantic space reflects human judgments we computed Spearman correlations between the predictions of the models and the human ratings. When calculating the correlations we included only those pairs of words that were present in the combined lexicon of the semantic spaces.

Similarity/Relatedness ratings - results

The average correlation between the Wordsim-353 subset of related words (n = 238) and the semantic distance measures derived from the count model was -0.35 (SD = 0.09; see Figure 4.). The negative correlations in this section reflect the fact that the relatedness measures are expressed in terms of distances rather than similarities.

For CBOW models it was -0.66 (SD = 0.04), and for the skip-gram model -0.59 (SD = 0.04). The measures derived from the count model correlated -0.43 (SD = 0.15) with the similarity subset of Wordsim-353 (n = 196), against -0.76 (SD = 0.02) for CBOW and -0.70 (SD = 0.04) for skip-gram.

The correlations with the Simlex-999 dataset (n = 998) were much lower. For the count model the average correlation was -0.10 (SD = 0.11), for CBOW it was -0.35 (SD = 0.06), and for skip-gram -0.27 (SD = 0.08).

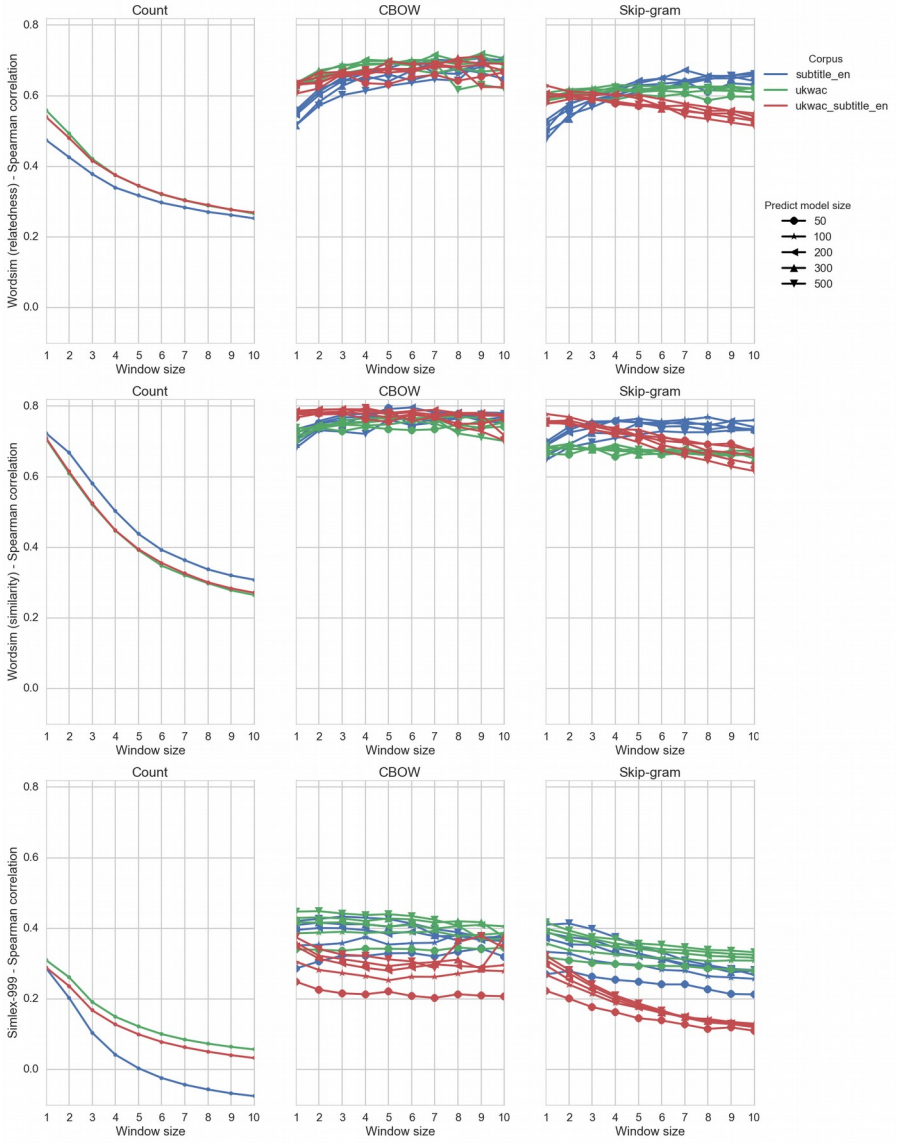

Figure 4. Performance of the three types of models on the similarity and relatedness ratings datasets (absolute values of correlations). There is a robust advantage of the predict models. The models trained on subtitle corpora underperformed compared to models trained on UKWAC or on the concatenation of the two corpora.

As shown in Figures 1-3, the worse performance of the count models was caused by the prediction power rapidly decreasing with

window size. The best performing count models had window size 1. For the Wordsim-353 relatedness subset and Simlex-999 the best count model was trained on the UKWAC corpus and correlated -0.55 and -0.31 with the human norms respectively. The best count model for the Wordsim-353 similarity subset (r = -0.72) also used window of size 1 but was trained on the subtitle corpus.

The skip-gram model also showed deteriorating performance with increasing window size for the similarity judgments. The best performing skip-gram models were the combined corpus with 100 dimensions and window size 1 for the Wordsim-353 similarity subset (r = -0.78), and on the UKWAC corpus with 500 dimensions and window size 1 for the Simlex-999 dataset (r = -0.42). Surprisingly, window size had a different effect for the relatedness judgments. For the Wordsim-353 relatedness subset, the best performance for a skip-gram model was achieved by training on the subtitle corpus with 200 dimensions and window size 7 (r = -0.67).

The CBOW models outperformed the other two model types. The best CBOW model for the Wordsim-353 similarity subset was trained on the subtitle corpus with 200 dimensions and window size 6 (r = -0.80), on the UKWAC corpus with 500 dimensions and window size 2 (r = -0.45) for the Simlex-999 dataset. For the Wordsim-353 relatedness subset, the best CBOW model was based on the UKWAC corpus with 200 dimensions and window size 9 (r = -0.72). Importantly, there was little effect of window size for the CBOW models (except for the smallest sizes, which resulted in less good performance).

Interestingly, for this task, models trained on individual corpora tended to perform better than models trained on the combination of corpora.

TOEFL - method

TOEFL is a dataset of 80 multiple choice questions created by linguists to measure English vocabulary knowledge in non-native speakers. The task of the person taking the test is to decide which of four candidate words is most similar to the target word. Landauer and Dumais (1997) first used this task to evaluate a distributional semantics model.

In our evaluation, we consider that a model provides a correct answer to TOEFL question when the correct candidate word has the smallest cosine distance to the target word in the semantic space compared to the other three candidate words. One point is awarded for that question in this case; zero points are given otherwise. When the target word or none of the four alternatives were present in the semantic space, we assigned a score of 0.25 to the item to simulate guessing.

TOEFL - results

The best count model (UKWAC corpus; window size 1) obtained a score of 83.7% on the TOEFL test. Average performance of the count models on this test was 61.2% (*SD* = 9.76%; see Figure 3.).

The predict model with the highest score on TOEFL was a CBOW model with 500 dimensions and window size 1, trained on the concatenation of the UKWAC and the subtitle corpora (score = 91.2%). The top skip-gram model was trained on the same corpus using the same window size but had a size of 300. On average the

CBOW models achieved a score of 73.4% (*SD* = 10.9%) and the skip-gram models a score of 69.0% (*SD* = 9.6%)

As shown in Figures 1-3, models trained on the subtitle corpora clearly performed worse on the TOEFL test than those trained on the UKWAC corpus or on the concatenation of both corpora. Like before, the count model showed a strong decrease in precision with increasing window size.

With a score of 87.5%, the semantic space from Baroni et al. (2014) surpassed the vast majority of our models.

DUTCH

Text corpus

We used the SONAR-500 text corpus (Oostdijk, Reynaert, Hoste, & van den Heuvel, 2013) and a corpus of movie subtitles to train the distributional semantic models.

The SONAR-500 corpus is a 500 million words corpus of contemporary Dutch and includes a wide variety of text types. It is aimed at providing a balanced sample of standard Dutch based on textual materials from traditional sources such as books, magazines and newspapers, as well as Internet based sources (Wikipedia, websites, etc.).

Tokens from the SONAR-500 corpus were extracted using the FoLIa toolkit[12]. We found that the corpus contained a small number of duplicate documents. In order to remove them from the corpus we ran the MinHash duplicate detection using *duometer* within each category of texts in the corpus. The final version of the SONAR-500 corpus,

[12] http://proycon.github.io/folia/

after duplicate detection and applying our tokenization procedure included 406 million tokens (1.9 million documents).

In order to compile the subtitle corpus, we downloaded 52,209 subtitle files. The corpus was cleaned in the same way as the English subtitle corpus. The final Dutch subtitle corpus contained about 26,618 documents and 130 million tokens.

Finally, we combined the SONAR-500 corpus and the subtitle corpus. As the SONAR-500 corpus also includes movie subtitles, we only included documents from the subtitle corpus that did not have a duplicate in the SONAR-500 corpus. This resulted in a combined corpus of 530 million tokens (1.926 million documents).

Model training

We used the same procedure for training the semantic spaces as the one used for the English corpora. For the Dutch material, we only used the models with window sizes of 1, 2, 3, 5 and 10, because our experience with the evaluation of the English semantic spaces had shown that the results vary most between the initial values and the general trend in performance is similar at higher window sizes.

When training the HAL-type model, 300,000 types with the highest frequency were used as word and contexts. The PPMI weighting scheme was applied to the resulting co-occurrence matrix. The same parameter settings as for English were applied when training the predict models. However, we trained only models with 200 and 300 dimensions.

Evaluation Tasks

Semantic priming - method

Because there is no large, publicly available dataset of semantic priming in Dutch, our analysis was limited to two smaller datasets. The first one was based on a lexical decision experiment conducted by Heyman, Van Rensbergen, Storms, Hutchison and De Deyne (2015), which included 120 target words, each preceded by related and unrelated words. We used only words from the low memory load condition and for each prime-target pair we used an average reaction time for the two SOAs (1200 and 200 ms) used in the experiment. This resulted in a dataset of 240 prime-target pairs with associated RTs. For 236 of these pairs both the prime and the target were present in our semantic spaces and were included in further analyses.

The second dataset on which we based our analysis was collected by Drieghe and Brysbaert (2002). This dataset includes 21 target words with one semantically related prime and two unrelated primes (one that was homophonic to the related prime and one that was completely unrelated). The small number of items in the second Dutch semantic priming dataset enabled only a very simple evaluation. In order to calculate how well each of the trained models fit the dataset we computed the distances between the primes and the targets for the related and the unrelated conditions, and we performed *t*-tests to verify whether the distances in the unrelated conditions were larger than in the related condition, as is the case for the human reaction times.

Semantic priming - results

In the dataset from Heyman et al. (2015) the baseline model including log of word frequency and length for both the prime and the target explained 4.82% of the variance in reaction times. An average performance of the models including various semantic predictors is presented in Table 2.

A conservative comparison of the baseline model with the model including relatedness measures derived from the worst performing semantic relatedness measure (count model based on the subtitle corpus trained with window size 10, 10.73% of variance explained) showed a highly significant contribution of this semantic predictor [$F(1, 230) = 11.05$, $p = 0.001$)].

On average, the models including the lexical predictors and semantic relatedness derived from the count models explained 13.61% of the variance in reaction times. The models including semantic relatedness derived from the skip-gram models explained 17.73% of the variance and the semantic predictors based on CBOW explained 19.05% of the variance. The best performing count model explained 16.20% of the variance in reaction times, and was trained on a concatenation of the subtitle and SONAR corpora with window size 2. The best skip-gram model explained 20.68% of the variance. That model had 200 dimensions and was trained on the concatenation of the two corpora using window size 5. The best CBOW relatedness measures, which explained 22.39% of the variance in RTs, had 200 dimensions and was trained on the concatenation of the two corpora using window size 10.

Table 2. The table shows the average results obtained from different classes of models for the words in different conditions in the two Dutch semantic priming experiments (Heyman et al., 2015; Drieghe & Brysbaert, 2002).

			Drieghe et al.					
		Heyman et al.	Cohen's d			Average distance		
Corpus	*Model*	*R^2*	*Related vs Control 1*	*Related vs Control 2*	*Control 1 vs Control 2*	*Related*	*Control 1*	*Control 2*
SONAR-500	HAL w. < 5	.148	.92	.98	.16	.91 (SD=.05)	.95 (SD=.02)	.95 (SD=.02)
	HAL w. >= 5	.119	.95	1	.15	.91 (SD=.05)	.95 (SD=.02)	.95 (SD=.02)
	CBOW	.191	.73	.73	-.03	.92 (SD=.06)	.95 (SD=.01)	.96 (SD=.01)
	skip-gram	.183	.56	.82	.45	.81 (SD=.09)	.85 (SD=.05)	.87 (SD=.04)
SONAR-500 + subtitle-nl	HAL w. < 5	.152	.55	.83	.48	.8 (SD=.1)	.84 (SD=.05)	.86 (SD=.04)
	HAL w. >= 5	.125	.43	.62	.4	.84 (SD=.09)	.87 (SD=.04)	.89 (SD=.03)
	CBOW	.207	1.34	1.25	-.12	.63 (SD=.16)	.85 (SD=.08)	.84 (SD=.1)
	skip-gram	.194	1.44	1.38	-.08	.57 (SD=.15)	.81 (SD=.08)	.81 (SD=.1)
subtitle-nl	HAL w. < 5	.140	1.36	1.15	-.38	.47 (SD=.19)	.77 (SD=.12)	.72 (SD=.16)
	HAL w. >= 5	.117	1.35	1.23	-.24	.48 (SD=.14)	.68 (SD=.08)	.66 (SD=.09)
	CBOW	.172	1.42	1.37	-.11	.38 (SD=.11)	.58 (SD=.08)	.57 (SD=.09)
	skip-gram	.153	1.34	1.09	-.34	.31 (SD=.13)	.52 (SD=.11)	.48 (SD=.14)

Note. The first column lists corpora on which the models were trained. The second column shows the different types of models. The HAL models using window sizes smaller and larger or equal to 5 are shown separately. Window sizes mattered less for the predict models so they are all reported together. The next column reports the percentage of variance explained in the dataset from Heyman et al. (2015). The following three columns display average effect sizes of comparisons between various conditions in the dataset from Drieghe & Brysbaert (2002). The last three columns report the mean and standard deviation between cues and targets in each of the conditions. All statistics are averaged over all parameter settings used to train the models.

For the Bayesian analysis we followed the same procedure as in the case of the English data. We first identified the best model based on lexical variables only. The analysis indicated that the best model included the logarithm of prime and target word frequency, and was strongly supported relative to a model including intercept only (BF_{10} = 29.01). We used this model as the reference in further analyses. Consideration of the subset of all models including lexical predictors and the worst performing semantic measure provided decisive evidence in favor of the model including the relatedness measures (in addition to log of prime word frequency; reference of the model based on lexical predictors only BF_{10} = 109.70).

When we ran a Bayes factor regression including the lexical predictors and the best performing semantic relatedness measures (CBOW model with 200 dimensions trained on the concatenation of the SONAR-500 and the subtitle corpora using window size 10), we found that the best model, overwhelmingly supported relative to the model based on lexical variables only (BF_{10} = 197,283,867), included the logarithm of prime and target word frequency in addition to the semantic relatedness measure.

In a direct comparison of the relatedness measures derived from each type of models (count, CBOW and skip-gram), the Bayes factor analysis indicated a decisive advantage of the model including relatedness measures derived from the best CBOW model, relative to the model including the best count relatedness measures (BF_{10} = 1682.73) and substantial evidence in favor of the CBOW relatedness measures relative to those derived from the skip-gram model (BF_{10} = 8.44). The best skip-gram relatedness measures were also decisively better than the best count relatedness measures (BF_{10} = 199.28).

The dataset from Drieghe and Brysbaert (2002) contained a set of target words with one related prime and two unrelated primes. Because the dataset was too small to run analyses at the item level, we limited ourselves to *t*-tests. Table 2 gives the average similarity scores for the various models. It clearly shows that the semantic relatedness was larger in the related condition than in the unrelated condition for all predict models. The situation was less convincing for the count models.

All predict models correctly simulated the expected pattern of results and showed that there was a significant difference between the related and the first unrelated condition (average *t*-test value = -6.08, *df* = 20, *SD* = 0.82; average *d* = 1.38, *SD* = 0.11; for all models $p < 0.001$ and effect sizes of $d > 1$) and between the related and the second unrelated condition (average *t*-test value = -5.16, *SD* = 1.10; average *d* = 1.25, *SD* = 0.17; for all models $p < 0.01$ and $d > 0.8$). Furthermore, there was no significant difference between the two unrelated conditions (average *t*-test value = 0.68; *SD* = 0.58; average *d* = 0.21, *SD* = 0.18; for all models $p > 0.1$ and $d < 0.5$).

For the count models the differences between conditions were much smaller. Using a significance level of $p < 0.05$, we obtained significance for only 11 out of 15 models between the related and the first unrelated condition (average *t*-test value = -2.53, *SD* = 0.82; average *d* = 0.73, *SD* = 0.22), and for only 14 out of 15 between the related and the second unrelated condition (average *t*-test value = -2.97; *SD* = 1.10; average *d* = 0.84, *SD* = 0.14). All the count models correctly showed no difference between the two unrelated conditions (average *t*-test value = -0.76; *SD* = 0.58; average *d* = 0.23, *SD* = 0.24; for all models $p > 0.10$). As could be expected on the basis of Figure

1, the count models with small window sizes did better than those with large window sizes.

Association norms - method

We used word association data from de Deyne and Storms (2008), who reported the associates most frequently given to 1,424 cue words. Like in the evaluation of the English data, we computed the average relative entropy between the probability distributions of the associates produced by our models and the human data.

Association norms - results

For the 1,424 cue words from de Deyne and Storms (2008), the baseline relative entropy score based on 10 randomly generated semantic vectors was 0.86 (*SD* = 0.0005; lower is better).

The average relative entropy for the count models was 0.78 (*SD* = 0.01). The best performing count model had a window size of 3 (trained on the SONAR-500 corpus), resulting in a relative entropy of 0.76.

The average relative entropy for the CBOW models was 0.79 (*SD*=0.03). The best performing model (relative entropy = 0.74) was trained on the combined SONAR-500 and subtitle corpus, had 200 dimensions and a window of size 10.

The average relative entropy for the skip-gram models was 0.80 (*SD* = 0.02) and the best performing model had the same parameters as the best performing CBOW model (relative entropy = 0.75).

INFLUENCE OF THE WINDOW SIZE

Our analyses indicated that the size of the window used to train the count models is an extremely important parameter when training these models. At the same time, it has to be acknowledged that the count and the predict models use the window size parameter differently during training. While the typical count models consider full window size for each target word, the predict models randomly choose a number between 1 and the requested window size and use that randomly chosen number for each single training case. This allows these models to utilize information about distant words but at the same time an average window size is reduced by half in such a procedure. To verify whether this aspect of the training can be responsible for the sharp drop in the performance of the count models that was not observed for the predict models we decided to train an additional set of count models using window sizes 1, 2, 3, 5, 7 and 10, on the English subtitle corpus and its concatenation with the UKWAC corpus. However, for this analysis we applied an analogous procedure of randomly choosing window size in each training step as it is the case for the predict models. As could be expected, we observed that using a randomized window size for training the count spaces decreased the speed at which performance of the spaces drops with increasing window size. Nevertheless, the performance was still best at window size 3, even when randomized window size was used. The improvement of using reduced window sizes was largest for the largest window sizes – for window size 10 the amount of explained variance increased by 0.7% (subtitle corpus) and 0.6% (concatenation of the text corpora) in LDT and by 0.1% (both subtitle corpus and the

concatenation of the two corpora) for naming. In naming, these improvements were comparable for window size 5 and 7, as well as window size 3 for the models trained on a concatenation of the corpora. In LDT, the improvement over the default models dropped by about 0.05% for window size 7 and then by another 0.2% for window size 5. In both tasks the changes for smaller window sizes were less than 0.1%.

This analysis indicates that the random reduction of the window size attenuates the decreasing performance of the count models, making them more comparable to the predict models even for larger window sizes. However, the general trend of optimal performance with a window size of about 3 can still be observed.

DISCUSSION

In this article we compared the performance of the recently proposed predict models of semantic similarity to the methods currently used in psycholinguistics by looking at how much variance the estimates explain in human performance data. In all cases, we saw an outcome that was at least equal to the existing measures and that was often superior to them. This was even true when we compared the measures based on semantic spaces to measures produced by human participants (e.g., word association norms or semantic features generated by participants), showing that the semantic vectors should be included in psycholinguistic research.

In line with previous findings (Baroni et al., 2014; Levy & Goldberg, 2014), the predict models were generally superior to the count models, although the best count models tended to come quite close to the predict models (and in a few cases even exceeded them).

The most important variable for the count models was window size, as shown by Bullinaria and Levy (2007, 2012). A problem in this respect, unfortunately, is that the optimal window size seems to depend on the task. It equals 3 for semantic priming, 1 for semantic relatedness judgments, and 2 for the prediction of word associations. The performance rapidly drops for non-optimal window sizes, as shown in Figure 1. At the same time, our additional analysis indicated that applying the same procedure of randomly selecting window sizes, as done in the predict models, is a way to attenuate the decrease in performance for larger window sizes.

In contrast, the predict models are less influenced by window size. In addition, their performance generally increases with window size (certainly up to 5). Of these models, the CBOW models typically outperformed the skip-gram models and there are no indications in the data we looked at to prefer the latter over the former. In general, there was little gain when the dimensions of the CBOW model exceeded 300 (sometimes performance even started to decrease; this was particularly true for semantic priming and word associations).

Given the superior performance of the CBOW models, it is important to understand the mechanisms underlying them. As a practical example of the CBOW model, we discuss the model that had the best average performance for English and that we also recommend for general use in psycholingusitic research (see also the section on availability below). This model is trained on the combined UKWAC and subtitle corpus, has a window size of 6, and contains 300 dimensions. There are input and output nodes for each word form in the corpus that is encountered at least 5 times, leading to about 904 thousand input and output nodes. The dimensionality of the model is

equal to the number of hidden nodes, which in this case is 300. The training of the model consists of the activation of the input nodes of the 6 words before the target word and the 6 words after the target word and predicting the activation of the output node corresponding to the target node. Over successive runs, the weights are adapted to improve performance. The semantic vector for a word consists of the 300 weights between the input node of a word and the hidden nodes after learning.

As shown in Figure 1, the CBOW model learns to predict the relationship between the target word and all words in the surrounding window simultaneously. In the HAL-type count model and in the skip-gram model, the relationship between the target word and each word in the window is trained individually. As a metaphor, consider a paper with a long set of co-authors of which one has been removed. The task is to predict the missing author. The HAL-type count model and the skip-gram model can only predict the missing author based on the individual co-occurrence between each known co-author and their past co-authors, which could result in the predicted co-author being completely unrelated to the other co-authors on the paper. The CBOW model, on the other hand, would predict the missing author based on the simultaneous consideration of all other co-authors on the paper. The model would be more likely to predict a co-author who often writes together with all or part of the co-authors than someone who frequently co-authors with only one of them.

In light of the current findings, it is important to understand the differences between the discussed models in Marr's (1982) terms. The count model specifies a computational problem for the cognitive system (learning to associate semantically related words) and provides

an abstract computational method for solving it using weighting schemes and dimensionality reduction. It has been argued (Levy & Goldberg, 2014) that the results of the skip-gram model can also be achieved by a certain type of a count model (PMI weighting shifted by a constant and dimensionality reduction steps) making the skip-gram model computationally equivalent to a count model. However, because the skip-gram model can be specified using prediction-based incremental learning principles in a neural network, it solves the computational problem posed by the count models in a way that is to a large extent psychologically plausible. Finally, although the CBOW model shares this algorithmic-level plausibility with the skip-gram model, CBOW cannot be reduced to a count model (Levy et al., 2015). Since the CBOW model compares favorably to the other investigated models it is an important task for future research to better understand this model at the computational level.

In this paper, we gave considerable attention to the type of corpus used to train a model. In computational linguistics, models are often found to perform best when trained on very large corpora (Banko & Brill, 2001) and this implies that register is second to size. Our data show that the large corpora typically used in computational linguistics are good for vocabulary tests, such as TOEFL but perform less well for psycholinguistic benchmarks such as semantic priming or word associate generation. On these tasks, corpora based on subtitles of films and television series perform better. When we consider what the TOEFL test requires, it is not surprising that training on very large corpora containing a large amount of specialist material is beneficial. Because TOEFL includes a large number of uncommon words, models trained on subtitle corpora can be expected to perform worse on this

test. Indeed, we would expect a person reading the material included in the very large corpus to score quite highly on the TOEFL and we would be equally unsurprised if a person watching only films and television series would perform worse. In contrast, the impressive performance of the relatively small corpora of subtitles on the semantic priming and word association tasks is surprising. This implies that when it comes to accounting for human behavior it is important to train models on a corpus that has a register closer to what humans experience. Recall that the TOEFL benchmark is not about predicting how well humans do, but about scoring as highly as possible. Associations in the larger corpus better reflect the semantic system for someone who scores very well on the TOEFL, whereas associations based on the subtitle corpus reflect more of a central tendency: As an example, our reference CBOW model based on subtitles, for *elephant* generates *giraffe*, *tusk*, *zoo,* and *hippo* as nearest semantic neighbors*;* on the other hand, the model trained on the combined UKWAC and subtitles corpus generates *howdah, tusked, rhinoceros,* and *mahout.* The first and second authors of this paper confess that they did not know what to make of two of the latter associations until they learned that a *howdah* is a seat for riding on the back of an elephant and that a *mahout* is a professional elephant rider. The example clearly illustrates how the models based on the larger corpora score higher on the TOEFL. Future research could investigate whether the advantage of the larger corpora is still maintained when the actual human responses are the benchmark instead of the highest score.

On the basis of the current study, conclusions about the relation between corpus register and size and human performance are

risky because these variables were not independent of each other. Still, it seems possible to conclude that given that the subtitle corpora are smaller *and* in many cases perform better on predicting semantic priming, the register of the subtitles better represents the input of human participants. On the other hand, the question remains what precisely in the bigger corpora accounts for the worse performance. Even adding subtitle material to the large corpora does not result in models that predict semantic priming as well as the subtitle corpora do alone. An answer may be that the smaller subtitle corpora result in close semantic relationships that are shared by many participants, while the large corpora result in more specialized semantic relationships that are known by only a few participants. This additionally suggest that increasing the size of a subtitle corpus further may not necessarily result in better performance on a semantic priming task because more specialized semantic relationships could be developed at the expense of more universally shared ones. This point is given further weight by taking into account that corpora over a certain size stop being ecologically realistic.[13]

Given the current set of results, we can unequivocally assert that distributional semantics can successfully explain semantic priming data, dispelling earlier claims (Hutchison et al., 2008). While Günther et al. (2015) found small effects for German, we obtain a strong and robust increase in the predictive power when the regression analysis includes semantic information derived from distributional semantics models. According to our analyses the predictions based on

[13]Assuming a maximum reading rate of 300 words per minute (Carver, 1989; Lewandowski, Codding, Kleinmann, & Tucker, 2003), a person who has read 16 hours per day for 18 years, has come across 300*60*16*365.25*18 = 1.89 billion words at most.

the semantic space models can match or exceed the ones based on human association datasets or feature norms. This is fortunate, because semantic similarity measures based on semantic spaces are available for many more words than human similarity or relatedness ratings and can be collected more easily for languages that do not yet have human ratings. In this regard, we should also point to recent advances in human data collection. For instance, collecting more than one response for each cue word in a word association task may lead to a more refined semantic network than the one we tested (De Deyne, Verheyen, Storms, 2015). It will be interesting to see how such a dataset compares to the semantic vectors we calculated.

Finally, it is of practical importance to mention that, at least for the semantic priming data, the pioneering LSA space available through a web-interface at the University of Colorado Boulder (1997) does not perform better than the reference semantic spaces we are releasing with the current paper. At the same time, it is surprising that the difference in performance is so small if we consider the size of the corpus (11 million words) on which the venerable LSA space was based. The relative success of LSA based on the small TASA corpus suggests that books used in schools are another interesting source of input (arguably because it is a common denominator. These books and subjects are read by most students).

AVAILABILITY

A big obstacle to the widespread use of distributional semantics in psycholinguistics has been the gap between the producers and potential consumers of such spaces. Although several packages have been published that allow users to train various kinds of semantic

spaces (e.g. S-Space, Jurgens & Stevens, 2010; DISSECT, Dinu, Pham, & Baroni, 2013; LMOSS, Recchia & Jones, 2009; HIDEX, Shaoul & Westburry, 2006), the large corpora and computational infrastructure as well as the technical know-how regarding training and evaluating semantic spaces is not available to many psycholinguists. Therefore, in order to encourage the exchange and use of semantic spaces trained by various research groups, we release a simple interface that can be used to measure relatedness between words on the basis of semantic spaces. Importantly, it can be used both as a standalone program and as a web-server that makes the semantic spaces available over the Internet. We believe that such an open-source contribution complements the existing ecosystem allowing researchers to train and explore semantic spaces (e.g. LSAfun; Günther, Dudschig, & Kaup, 2014). We encourage contribution from other researchers to the code base for our interface, which is hosted on a platform for sharing and collaborative development of programming projects.[14]

To make it as easy as possible for the authors of semantic spaces to work with our interface, two simple formats are used: the Character Separated Values (CSV) format and the matrix market format[15] that supports efficient representations of sparse matrices such as those created when training count models without dimensionality reduction.

We release a series of predict and count spaces for Dutch and English that were found to be consistently well performing in the

[14] The code is available at the address: http://crr.ugent.be/snaut/

[15] For more information about the matrix market format see: http://math.nist.gov/MatrixMarket/

present evaluations. Each of the spaces is released in a format compatible with our interface. The predict spaces can be also used with the LSAfun (Günther et al., 2014) interface.

In addition to the full semantic spaces for English and Dutch used for the present study we also make available smaller subspaces which may be very useful in many cases, as they can be explored using very limited computational resources. The smaller semantic spaces are based on two subset tokens from full space:

1. a subset of the 150,000 most frequent words in each of the spaces
2. a subset based on the lemmas found in the corpora

Information about how well each of the released semantic spaces performed on our evaluation tasks is shown in Tables 3 (for English) and 4 (for Dutch).

As semantic spaces can always be improved by finding superior methods or parameter settings, we know that the spaces that we trained can and will be outperformed by other spaces. Our interface fully encourages such developments.

Table 3. The performance of the released English semantic spaces on the evaluation tasks.

			Semantic priming project											
			Lexical decision		*Naming*		*Associations relative entropy*	*Simlex-999*		*Wordsim-353 relatedness*		*Wordsim-353 similarity*		*TOEFL*
subset	*model*	*N*	*R^2 baseline*	*R^2 model*	*R^2 baseline*	*R^2 model*		*N*	*r*	*N*	*r*	*N*	*r*	*score*
lemmas	subtitle, CBOW, dim. 300, window 6	5311	.399	.465	.319	.337	.696	999	-.414	236	-.672	196	-.765	.559
top 150000	subtitle, CBOW, dim. 300, window 6	5738	.389	.455	.312	.331	.698	998	-.412	238	-.671	196	-.765	.663
full	subtitle, CBOW, dim. 300, window 6	5738	.389	.455	.312	.331	.698	999	-.414	238	-.671	196	-.765	.663
lemmas	subtitle, count, window 3	5311	.399	.471	.319	.339	.696	999	-.106	236	-.382	196	-.581	.494
top 150000	subtitle, count, window 3	5738	.389	.457	.312	.332	.699	998	-.104	238	-.378	196	-.581	.663
top 300000	subtitle, count, window 3	5738	.389	.457	.312	.332	.699	999	-.106	238	-.378	196	-.581	.659
lemmas	UKWAC + subtitle, CBOW, dim. 300, window 6	5311	.399	.454	.319	.338	.633	999	-.301	236	-.673	196	-.776	.666
top 150000	UKWAC + subtitle, CBOW, dim. 300, window 6	5738	.389	.445	.312	.331	.636	998	-.3	238	-.676	196	-.776	.834
full	UKWAC + subtitle, CBOW, dim. 300, window 6	5738	.389	.445	.312	.331	.636	999	-.301	238	-.676	196	-.776	.853
lemmas	UKWAC + subtitle, count, window 1	5311	.399	.458	.319	.336	.708	998	-.289	236	-.54	196	-.71	.628
top 150000	UKWAC + subtitle, count, window 1	5738	.389	.448	.312	.330	.712	998	-.289	238	-.54	196	-.71	.809
top 300000	UKWAC + subtitle, count, window 1	5738	.389	.448	.312	.330	.712	998	-.289	238	-.54	196	-.71	.828

Table 4. The performance of the released Dutch semantic spaces on the evaluation tasks. For the evaluation based on data from Heyman et al. (2015) all datasets included 236 prime-target pairs and the baseline model based on lexical predictors explained 6.44% of the variance in RTs. The only exception were models based on the 150,000 most frequent words which included 264 prime-target pairs (baseline model explained variance: 6.22%). All models based on Drieghe et al. 2015 included 63 pairs of words.

			Drieghe et al.						
		Heyman et al.	Cohen's d			Average distance			Associations relative entropy
subset	model	R^2	Related vs Control 1	Related vs Control 2	Control 1 vs Control 2	Related	Control 1	Control 2	
lemmas	SONAR-500, count, window 3	.143	.832	.976	.334	.883 (SD=.062)	.925 (SD=.027)	.934 (SD=.024)	.766
top 150000	SONAR-500, count, window 3	.143	.832	.976	.334	.883 (SD=.062)	.925 (SD=.027)	.934 (SD=.024)	.764
top 300000	SONAR-500, count, window 3	.143	.832	.976	.334	.883 (SD=.062)	.925 (SD=.027)	.934 (SD=.024)	.765
lemmas	SONAR-500 + subtitle, count, window 2	.162	.991	1.03	.156	.905 (SD=.050)	.947 (SD=.017)	.950 (SD=.020)	.774
top 150000	SONAR-500 + subtitle, count, window 2	.162	.991	1.03	.156	.905 (SD=.050)	.947 (SD=.017)	.950 (SD=.020)	.772
top 300000	SONAR-500 + subtitle, count, window 2	.162	.991	1.03	.156	.905 (SD=.050)	.947 (SD=.017)	.950 (SD=.020)	.773
lemmas	SONAR-500 + subtitle, CBOW, dim. 200, window 10	.224	1.532	1.542	.044	.633 (SD=.149)	.904 (SD=.069)	.907 (SD=.068)	.745
top 150000	SONAR-500 + subtitle, CBOW, dim. 200, window 10	.222	1.532	1.542	.044	.633 (SD=.149)	.904 (SD=.069)	.907 (SD=.068)	.739
full	SONAR-500 + subtitle, CBOW, dim. 200, window 10	.224	1.532	1.542	.044	.633 (SD=.149)	.904 (SD=.069)	.907 (SD=.068)	.743

ACKNOWLEDGEMENT

This research was made possible by an Odysseus grant from the Government of Flanders.

REFERENCES

Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118,* 438-482.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual Word Recognition of Single-Syllable Words. *Journal of Experimental Psychology: General*, *133*(2), 283–316. http://doi.org/10.1037/0096-3445.133.2.283

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: Large scale analysis of lexical processes. In J. S. Adelman (Ed.), *Visual Word Recognition Vol. 1: Models and Methods, Orthography and Phonology*. Hove, England: Psychology Press.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Vol. 1). Retrieved from http://clic.cimec.unitn.it/marco/publications/acl2014/baroni-etal-countpredict-acl2014.pdf

Banko, M., & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (pp. 26–33). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1073017

Broder, A. Z. (1997). On the resemblance and containment of documents. In *Compression and Complexity of Sequences 1997. Proceedings* (pp. 21–29). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=666900

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A.M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and

implications for the choice of frequency estimates in German. *Experimental Psychology, 58,* 412-424.

Brysbaert, M., Keuleers, E., & New, B. (2011). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology, 2*:27.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. http://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding Part-of-Speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods, 44*, 991-997.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510–526.

Bullinaria, J. A., & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior Research Methods*, *44*(3), 890–907.

Carver, R. P. (1989). Silent reading rates in grade equivalents. *Journal of Literacy Research, 21*(2), 155-166.

Chiarello et al. (1990). Semantic and associative priming in the cerebral hemispheres: some words do, some words don't ... sometimes, some places. - PubMed - NCBI. Retrieved April 27, 2015, from http://www.ncbi.nlm.nih.gov/pubmed/2302547

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Coltheart, M., Davelaar, E., Jonasson, J., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we?. *Psychological Bulletin, 74*(1), 68-80.

De Deyne, S., Verheyen, S., & Storms, G. (2015). The role of corpus size and syntax in deriving lexico-semantic representations for a wide range of concepts. *The*

Quarterly Journal of Experimental Psychology, *68*(8), 1643–1664. http://doi.org/10.1080/17470218.2014.994098

Dinu, G., Pham, N., Baroni, M. (2013). DISSECT: Distributional semantics composition toolkit. *Proceedings of the system demonstrations of ACL 2013 (51st Annual Meeting of the Association for Computational Linguistics)* (pp. 31–36). East Stroudsburg, PA, ACL.

Drieghe, D., & Brysbaert, M. (2002). Strategic effects in associative priming with words, homophones, and pseudohomophones. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(5), 951–961. http://doi.org/10.1037//0278-7393.28.5.951

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google* (pp. 47–54). Retrieved from http://www.researchgate.net/profile/Adam_Kilgarriff/publication/237127354_Proceedings_of_the_4_th_Web_as_Corpus_Workshop_(WAC4)_Can_we_beat_Google/links/00b7d5290647fbc33f000000.pdf#page=53

Günther, F., Dudschig, C., & Kaup, B. (2014). LSAfun - An R package for computations based on Latent Semantic Analysis. *Behavior Research Methods*. http://doi.org/10.3758/s13428-014-0529-0

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*, *69*(4), 626–653. http://doi.org/10.1080/17470218.2015.1038280

Harris, Z. (1954). Distributional structure. *Word, 10*(2-3), 1456–1162.

Heyman, T., Van Rensbergen, B., Storms, G., Hutchison, K. A., & De Deyne, S. (2015). The influence of working memory load on semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 911–920. http://doi.org/10.1037/xlm0000050

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences*, *11*(10), 428–434. http://doi.org/10.1016/j.tics.2007.09.004

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. http://doi.org/10.1126/science.1127647

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, *61*(7), 1036–1066. http://doi.org/10.1080/17470210701438111

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., … Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*(4), 1099–1114. http://doi.org/10.3758/s13428-012-0304-z

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552. http://doi.org/10.1016/j.jml.2006.07.003

Juhasz, B. J., Lai, Y.-H., & Woodcock, M. L. (2015). A database of 629 English compound words: ratings of familiarity, lexeme meaning dominance, semantic transparency, age of acquisition, imageability, and sensory experience. *Behavior Research Methods*, *47*(4), 1004–1019. http://doi.org/10.3758/s13428-014-0523-6

Jurgens, D., & Stevens, K. (2010). The S-Space package: an open source package for word space models. In *Proceedings of the ACL 2010 System Demonstrations* (pp. 30–35). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1858939

Kass, R.E., & Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Associatio,* 430 (90), 773-395.

Keuleers, E., & Balota, D. A. (2015). Megastudies, Crowdsourcing, and Large Datasets in Psycholinguistics: An Overview Of Recent Developments. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1457–68. http://doi.org/10.1080/17470218.2015.1051065

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. http://doi.org/10.3758/BRM.42.3.643

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice Effects in Large-Scale Visual Word Recognition Studies: A Lexical Decision Study on 14,000

Dutch Mono- and Disyllabic Words and Nonwords. *Frontiers in Psychology*, *1*. http://doi.org/10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2011). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. http://doi.org/10.3758/s13428-011-0118-4

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review; Psychological Review*, *104*(2), 211.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. http://doi.org/10.1038/nature14539

Levy, O., Goldberg, Y., Dagan, I., & Ramat-Gan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*. Retrieved from http://u.cs.biu.ac.il/~nlp/wp-content/uploads/Improving-Distributional-Similarity-TACL-2015.pdf

Lewandowski, L. J., Codding, R. S., Kleinmann, A. E., & Tucker, K. L. (2003). Assessment of reading rate in postsecondary students. Journal of Psychoeducational Assessment, 21(2), 134-144.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of *g* Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, *103*(481), 410–423. http://doi.org/10.1198/016214507000001337

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *The Quarterly Journal of Experimental Psychology*, 1–20. http://doi.org/10.1080/17470218.2014.988735

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2014). Subtlex-pl: subtitle-based word frequency estimates for Polish. *Behavior Research Methods*. http://doi.org/10.3758/s13428-014-0489-4

Marr, D. (1982), *Vision: A Computational Approach*, San Francisco: Freeman & Co.

McDonald, S., & Brew, C. (2004). A Distributional Model of Semantic Context Effects in Lexical Processing. In D. Scott, W. Daelemans, & M. A. Walker (Eds.), *ACL* (pp. 17–24). ACL. Retrieved from http://dblp.uni-trier.de/db/conf/acl/acl2004.html#McDonaldB04

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. Retrieved from http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Miller, R.R., Barnet, R.C., Grahame, N. J. (1995). Assessment of the Rescorla-Wagner Model. *Psychological Bulletin, 117* (3), 363-386.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://w3.usf.edu/FreeAssociation/.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(04). http://doi.org/10.1017/S014271640707035X

Olshausen, B. A., & Field, D. J. (1996). Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature*, *381*, 607–609.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199.

Patel et al. (1997). ESRLTC.ps. Retrieved April 27, 2015, from https://www.cs.bham.ac.uk/~jxb/PUBS/ESRLTC.ps

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*(3), 647–656. http://doi.org/10.3758/BRM.41.3.647

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The*

Quarterly Journal of Experimental Psychology, *68*(8), 1584–1598. http://doi.org/10.1080/17470218.2014.941296

Rehuřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research*, *47*(6), 877–903. http://doi.org/10.1080/00273171.2012.734737

Rumelhart, D. E., Hinton, G. E., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536. doi: 10.1038/323533a0

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, *38*(2), 190–195.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, *88*(2), 135.

Testolin, A., Stoianov, I., Sperduti, A., & Zorzi, M. (2015). Learning Orthographic Structure With Sequential Generative Neural Networks. *Cognitive Science*. Advance online publication. http://doi.org/10.1111/cogs.12258

Zorzi, M., Testolin, A., & Stoianov, I. P. (2013). Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Frontiers in Psychology*, *4*. http://doi.org/10.3389/fpsyg.2013.00515

Wagenmakers, E.-J. (2007). A practical solution to the pervasive provlems of p values. *Psychonomic Bulletin & Review*, 14 (5), 779-804.

Widrow, G., & Hoff, M. E. (1960). *Adaptive switching circuits.* Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, Part 4, 96-104

Chapter 6. How useful are corpus-based methods for extrapolating psycholinguistic variables?[1]

ABSTRACT

Subjective ratings for age of acquisition, concreteness, affective valence, and many other variables are an important element of psycholinguistic research. However, even for well-studied languages, ratings usually cover just a small part of the vocabulary. A possible solution involves using corpora to build a semantic similarity space and to apply machine learning techniques to extrapolate existing ratings to previously unrated words. We conduct a systematic comparison of two extrapolation techniques: k-nearest neighbours, and random forest, in combination with semantic spaces built using latent semantic analysis, topic model, a hyperspace analogue to language (HAL)-like model, and a skip-gram model. A variant of the k-nearest neighbours method used with skip-gram word vectors gives the most accurate predictions but the random forest method has an advantage of being able to easily incorporate additional predictors. We evaluate the usefulness of the methods by exploring how much of the human performance in a lexical decision task can be explained by extrapolated ratings for age of acquisition and how precisely we can assign words to discrete categories based on extrapolated ratings. We find that at least some of the extrapolation methods may introduce artefacts to the data and produce results that could lead to different conclusions that would be reached based on the human ratings. From a

[1] This chapter was published as Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based methods for extrapolating psycholinguistic variables? *Quarterly Journal of Experimental Psychology*, *68*(8), 1623-1642.

practical point of view, the usefulness of ratings extrapolated with the described methods may be limited.

INTRODUCTION

Human ratings for variables such as age of acquisition (AoA), imageability, concreteness, or affective ratings are an indispensable element of psycholinguistic research. They are also notoriously difficult to collect. Even though it is now possible to obtain measurements for tens of thousands of words more efficiently by using crowdsourcing platforms (Brysbaert, Warriner, & Kuperman, 2014; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Warriner, Kuperman, & Brysbaert, 2013), collecting human ratings for all words in all languages for all variables is a daunting task.

Potentially, this problem could be alleviated by supplementing traditionally collected ratings with extrapolated ratings. However, to make this possible, we need to identify methods for extrapolating rating data and find sources of information on which the predictions could be based. Some psycholinguistic variables have evident predictors. For instance, the strong correlation of word frequency with AoA (for a review see Brysbaert & Ghyselinck, 2006) makes word frequency one clear candidate predictor for this variable. However, frequency does not predict AoA completely, and as for other variables such as imageability, concreteness, or affective ratings it appears that predictors should also include semantic word properties. For instance, it would be much easier to predict the valence rating for the word “birthday” if we knew ratings for the words “cake” and “party”, assuming that the three words are semantically closely related. Even in the case of AoA, an inspection of available ratings suggests that semantics may bring substantial information to the prediction of this variable because, at least to some extent, words related to similar

topics are more likely to be acquired around the same age. For example, words related to family or food are often acquired early while words related to violent crime and disease are acquired later.

The idea of using semantic information to extrapolate ratings is not new. Sources of such information—for example, WordNets, databases in which lexemes are grouped into sets of synonyms and linked based on semantic and lexical relations—or co-occurrence models derived from text corpora have already been used to accomplish this task. For instance, Bestgen (2002) and Bestgen and Vincze (2012) proposed an extrapolation method based on semantic similarity of a target word to a number of rated words in a semantic space created using latent semantic analysis (LSA; Landauer & Dumais, 1997), taking their averaged rating as an extrapolated rating of the target word. The authors based their analyses on the ANEW (affective norms for English words) norms (Bradley & Lang, 1999) for valence, arousal, and dominance as well as on concreteness and imagery ratings collected by Gilhooly and Logie (1980). Their method turned out to produce high correlations for this set of norms. Along the same lines, Feng, Cai, Crossley, and McNamara (2011) proposed that semantic information obtained from WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990) and from LSA can be used together with data about other lexical properties to train a regression model and to predict human ratings of concreteness for 3521 nouns from the Medical Research Council (MRC) database (Coltheart, 1981). In a similar fashion, word co-occurrence information derived from a text corpus with High Dimensional Explorer (HiDEx) (Shaoul & Westbury, 2006, 2010) was used to estimate imageability (Westbury et al., 2013) and subjective familiarity (Westbury, 2013) ratings. In

addition, Recchia and Louwerse (2015) used Google Ngrams to train a hyperspace analogue to language (HAL)-like model and used them to predict affective ratings. They obtained even higher correlations between the original ratings and the reconstructed ratings when a linear model was used to combine the extrapolated ratings based on the semantic space and other psycholinguistic variables.

CURRENT STUDY

Although the results of previous studies show that word similarities derived from textual materials are an important source of information for extrapolating psycholinguistic ratings, details of the extrapolation procedures in these studies were too heterogeneous to allow for direct comparison of their efficiency; they used different sets of predictors, information derived from different corpora, different kinds of models, and different validation procedures. In addition, ratings are often used by researchers to split stimuli into groups rather than used as fully continuous variables. Therefore measuring the correlation with original ratings may be insufficient to fully evaluate the usefulness of the proposed methods for practical research purposes. Moreover, because correlations consider only standardized variables, they do not tell us anything about whether the extrapolation procedure preserves the scale that was used for measuring the original ratings and how close the extrapolated ratings are to the original ratings if the original scale were used. Finally, we have to ensure that the extrapolated variables are not contaminated by artefacts that may arise when the extrapolation methods are applied.

In the current paper we systematically evaluate and compare different extrapolation methods. We use very large datasets of

subjective ratings for English words, which allow us to evaluate how well extrapolation techniques work for tens of thousands of words. We investigate the quality of the predictions made by two different methods (k-nearest neighbours and a random forest) on the basis of four different models from which word similarities can be extracted: LSA, a method based on the HAL (Lund & Burgess, 1996), a topic model (Blei et al., 2003), and a recent skip-gram approach (Mikolov, Chen, Corrado, & Dean, 2013).

In addition to considering the correlations between the original and extrapolated ratings, we evaluate how useful the extrapolated ratings are for explaining human performance in a behavioural task. In order to evaluate whether the extrapolated variables can be used as a replacement of the original variables, we need to ensure that they have the same properties as the original ratings. We also evaluate the performance of ratings extrapolated with different methods compared to that of the original ratings when dichotomization and binning procedures are applied.

Unlike word association norms or WordNets, all predictors used in our analyses can be automatically derived from a text corpus. Such a choice of predictors is optimal if the primary goal of the applied methods is to make it possible to obtain predictions of ratings for different variables for words in many languages in which resources such as association norms or WordNets may not exist yet. Our primary analyses are also based on extrapolations with relatively small training sets to better simulate a situation in which only a limited set of rated words is available in a given language.

Representing similarity between words

LSA, topic models, HAL, and the skip-gram model are methods that make use of patterns of word co-occurrence in textual materials to reconstruct some of the semantic structure of a language. They are typically trained on large text corpora, and, although the details of the training procedures are fundamentally different, their results can be interpreted as vector representations of words in a continuous multidimensional space.

LSA (Landauer & Dumais, 1997) starts with a matrix with n rows representing words and k columns representing documents. A number in each cell of the matrix represents the count of occurrences of a particular word in a particular document. In the next step, singular value decomposition (SVD), a matrix decomposition technique from linear algebra, is applied to the matrix, reducing its dimensionality to a much smaller number m. If we think of each word as a point in a multidimensional space, the goal of applying this technique is to reduce the representation of a word from a point in a k-dimensional space to a point in an m-dimensional space while preserving most of the similarity structure between words. In other words, by applying this mathematical method one obtains a more compact representation than the full word by document matrix. A limitation of this method is that after the transformation the obtained dimensions do not correspond to interpretable topics.

Topic models are a set of probabilistic methods to discover thematic structure in a collection of documents. Latent Dirichlet allocation (LDA; Blei et al., 2003) is perhaps the most popular method based on this approach. For LDA each document in a text corpus is a mixture of topics, which, in turn, represent probability distributions

over words. LDA assumes that a text corpus is a product of a generative process, according to which each word in a document is generated by sampling a topic from a probability distribution over topics and then by sampling a word from the probability distribution of the words in the selected topic. In order to reverse this process and infer a probability distribution from a text corpus, one can apply methods such as Gibbs sampling (Geman & Geman, 1984) or variational inference (Jordan, Ghahramani, Jaakkola, & Saul, 1999). Describing the details of these methods is beyond the scope of this paper. What is important for our goals is that, based on LDA, one can obtain probability distributions of topics for each document and a probability distribution of words for each of the topics. In each topic a group of semantically related words obtains high probabilities. For instance, the method may discover a topic in which words such as birthday, happy, cake, party, day, gift, surprise, and love have high probabilities but semantically unrelated words have low probabilities. A second topic may include gun, shoot, kill, bullet, shot, fire, weapon with high probabilities, and so on. Although the default interpretation of LDA is expressed in probabilistic terms, it can also have a geometric interpretation (Steyvers & Griffiths, 2007), which is similar to that described in the case of LSA. The important difference between LSA and topic models is that the probability distributions produced by the latter method can be interpreted as corresponding to meaningful thematic groups. Because one of the results produced by the topic model is an assignment of all words in a text corpus to individual topics, in the current study we used vectors with the number of such assignments, normalized with word frequency, as a topic model representation of words.

Yet another approach to reconstructing semantic space from word co-occurrence is taken by the HAL model (Lund & Burgess,1996). In this approach the co-occurrences are collected by moving a window through the corpus. The window includes a certain number of words, and the number of times each pair of words co-occurs in a window is counted. By default, no dimensionality reduction technique is applied to the co-occurrence matrix, so resulting word vectors store many more values than in the case of LSA or topic models. This can be a problem if the resulting matrix is used as a basis for further processing. In this paper we use a HAL-like model in which co-occurrence counts are weighted with a positive pointwise mutual information (PMI; e.g., Recchia & Jones, 2009) scheme. In this approach the raw co-occurrence counts are substituted by a measure rooted in information theory, which can be computed according to the following formula:

$$PMI(x,y)=\log_2\frac{p(x,y)}{p(x)p(y)}$$

Where p(x,y) can be calculated as the ratio between number of co-occurrences of two words divided by the total number of words in the corpus, while p(x) and p(y) are the frequencies of each of the two words divided by the total number of words in the corpus. In the next step all negative values are removed from the matrix (Manning & Schütze, 1999).

It is important to note the difference between the bag-of-words approach used in LSA and topic models, and the approach taken by HAL: The former methods consider global, document-level co-occurrence patterns whereas HAL is based on local word co-occurrences within a relatively narrow window.

The fourth approach to modelling semantics that we consider was recently developed by Mikolov, Chen, et al. (2013), who proposed that word vectors can be efficiently computed by using skip-grams combined with a simple two-layer neural network. In this approach, a network is trained by presenting words from a corpus and trying to predict each of the words in a small window surrounding that word. The network uses a stochastic gradient descent computed using a back-propagation rule (Rumelhart, Hinton, & Williams, 1986) to learn from errors that it makes in its predictions and by adjusting weights in the network accordingly. When the training is finished, the weights of the connections in such a network are extracted and used as vector representations of words. Because similar words tend to occur in similar contexts, they tend to have similar vectors. Baroni, Dinu, and Kruszewski (2014) evaluated different types of models in a comprehensive set of tasks and found that models using methods that are based on predicting the context, as is the case for the skip-gram model, rather than on counting word co-occurrences tend to produce word vectors that better capture word similarities. Moreover, prediction-based approaches turned out to be more robust to different parameter choices.

Similarly to HAL, the skip-gram method is based on word co-occurrences in a narrow window rather than bag-of-words as is the case for LSA and topic models. An interesting contrast between bag-of-words models and models based on narrow windows is that the former are usually considered to be better at modelling thematic information and to outperform window-based methods in tasks such as predicting human associations, while window-based methods seem

to be better at modelling taxonomical relations, synonymy, or grammatical relationships (Sahlgren, 2006).

Since all four discussed semantic space models represent words as multidimensional vectors we can use them to extrapolate ratings in a very similar way.

Extrapolation methods

In this section, we describe the different extrapolation methods used in the current study. Specific parameter settings are reported in the Method section.

K-nearest neighbours

Bestgen and Vincze (2012) proposed that a variant of the k-nearest neighbours method (Fix & Hodges, 1951) can be used to extrapolate human ratings. According to this approach, for each word in the test set we identify the set of the most similar words (as measured with cosine distance) in the training set and assign the mean rating of these words to the target word as the extrapolated rating. The number of words that are considered in the averaging is a parameter of the model. For instance, according to the skip-gram model trained on our corpus, the five most similar words to gun are pistol, rifle, weapon, revolver, and shoot with corresponding arousal scores of 5.79, 6.14, 6.27, 6.29, and 6.00 in a set of norms published by Warriner et al. (2013). Assuming that the number of considered neighbours would be set to 5 and that all these words would be found in the training set, the model would predict that the arousal value for gun should be equal to the mean of the arousal values for these words (6.09).

Bestgen and Vincze (2012) investigated the optimal neighbourhood size that should be considered when predicting ratings and found that the accuracy of the method increased with neighbourhood size up to a value of 30.

Random forest

Random forest (Breiman, 2001) is a general-purpose machine-learning technique based on an ensemble of randomized decision trees. It builds a set of decision trees where each tree is based on a slightly different sample of the full dataset, reducing the risk of overfitting the model. Each decision tree is created by recursively splitting the dataset into smaller and smaller subsets in a way that maximizes information about the predicted variable. For instance, the method could potentially decide that a split at a certain value of a particular predictor (for instance, a topic with high probability of words such as birthday, happy, cake, party, day, gift, surprise, love) allows the full dataset to be divided into two subsets with more homogeneous valence in each of the two subsets than in the case of other splits. It would then try to further break each of the two subsets into smaller and smaller subsets, finally creating a decision tree, where at each step the decision about which branch to follow is made based on the value of a particular predictor. Then, in order to make a global prediction, the predictions of the individual trees contained in the model are averaged (in the case of a regression problem) or votes for different classifications are counted (in the case of a classification problem). The method has been shown to give accurate predictions in many different applications. Since the default parameter settings for random forests work well in a wide range of applications, the method can be considered as effectively nonparametric. The method is also

resistant to overfitting, even if a very large number of predictors is included in the model, making it well suited for our purpose. It allows us to use the values assigned to each word on all the individual dimensions of a word vector as separate predictors. Moreover, this set of predictors can be extended to include additional variables (both continuous and categorical). The drawback of the random forest method is that it makes it difficult to examine the exact relationship between the predictors and a predicted value.[2]

METHOD

MATERIALS

Ratings

To train and test the extrapolation methods, we used large sets of norms for multiple variables: AoA ratings for 30,121 words (Kuperman et al., 2012), concreteness ratings for 37,058 words (Brysbaert et al., 2014), and affective ratings (arousal, dominance, valence) for 13,915 words (Warriner et al., 2013).

The reliability of the ratings can be considered the upper bound for the performance of the extrapolation procedures. The split-half reliabilities, as reported in the respective publication, were equal to . 915 for AoA (Kuperman et al., 2012), .914 for valence, .689 for arousal, and .770 for dominance (Warriner et al., 2013). Concreteness ratings correlated .92 with the ratings in the MRC database, suggesting a high reliability for the dataset as well (Brysbaert et al., 2013).

[2] For random forest the function that describes the relationship does not have to be linear or even continuous.

Text corpus

Because subtitle corpora were shown to be particularly adequate for conducting psycholinguistic research (e.g., Brysbaert & New, 2009; Keuleers, Brysbaert, & New, 2010) and because subtitle corpora can be easily collected for many languages for which we may want to extrapolate ratings, the semantic spaces and word frequencies that were used in the current study were based on an English subtitle corpus including about 385 million words. To compile the corpus we downloaded 204,408 documents containing film and television subtitles flagged as English by the contributors of Open Subtitles website (http://opensubtitles.org) and removed all subtitle-specific text formatting before further processing. In order to remove documents containing large fragments of text in languages other than English, we calculated preliminary word frequencies and excluded all documents in cases where the 30 most frequent words did not cover at least 10% of the total number of tokens in that document. Because many documents are available in multiple versions, it was necessary to remove duplicates from the corpus. To do so, we used a custom method based on clustering documents with similar thematic structure derived from a topic model trained on all the files. If any pair of files within a cluster had an overlap of at least 10% unique word trigrams, we removed one of the files from the corpus. The resulting dataset contained 69,382 documents.[3]

Based on that corpus we calculated word frequencies for all word forms. We also lemmatized the corpus with the Stanford tagger (Toutanova, Klein, Manning, & Singer, 2003; Toutanova & Manning,

[3] We later compared the result of this procedure with a standard MinHash approach to removing near-duplicates (Broder, 1997). The resulting sets of files overlapped in 98.5%.

2000). Because the resulting set of part of speech tags was too complex for our purposes, we used a simplified set of tags (see Supplemental Material).

General approach

To systematically study the performance of the different prediction methods using word vectors obtained from models implementing different approaches to distributional semantics, we ran 10 iterations of the following cross-validation procedure for each of the variables:

1. We split the whole set of rated words into a test set and a training set. The results reported first are based on a split of the full datasets into training and test sets with 25% of the data in the training set and 75% of the data in the test set. Later in the paper, we also examine the influence of the size of the training set on the prediction accuracy.
2. Using the data from each of the word vector models, we trained a k-nearest neighbours and a random forest model using data in the training set and then extrapolated the ratings for the words in the test set. The only exception was the HAL-like model, for which, because the large number of dimensions made the problem too computationally demanding for the random forest, we were able to train only the k-nearest neighbours model. As a baseline, we also trained three linear models with the following sets of predictors: (a) $\log_{10}$ of word frequency as the only predictor, (b) $\log_{10}$ of word frequency, word length (number of letters), and a measure of orthographic neighbourhood density (OLD20; Balota et al., 2007), and (c) a model including the same predictors as those in the second model plus a measure of semantic

neighbourhood density (inverse N count; Shaoul & Westbury, 2006, 2010). The baseline linear models did not include information obtained from the semantic spaces.

3. We evaluated the performance of the method by correlating the predicted ratings with the original ratings in the test set.

We decided to use this approach as it clearly indicates the predictive accuracy of the models and allows us to draw conclusions that avoid the risk of being based on overfitting. The results of the 10 iterations can be compared to those for baseline models, based on identical sets of words in the training and test sets.

Semantic spaces

Because the norms that were used to train and validate the extrapolation procedure were mostly ratings of lemmas, we also used a lemmatized text corpus (with base forms in place of inflected forms) to train the semantic models.

Following a common practice, in the case of bag-of-words models (LSA and a topic model) we removed very frequent and very rare words from the corpus before training. The lemmas in the high-frequency stop-list included about 500 common English words. As in the procedure applied by Bestgen and Vincze (2012), words occurring in the corpus fewer than 10 times were removed as well.

When creating the LSA model, prior to submitting the document-term matrix to SVD, we applied a term-frequency times inverse document-frequency transformation, which is a common weighting scheme used in information retrieval (e.g., Manning, Raghavan, & Schütze, 2008).

To preserve the same dimensionality for all the methods involving some form of dimensionality reduction, we used 600 eigenvectors corresponding to the highest singular values in the LSA model, 600 topics in the LDA topic model, and a 600 dimensional skip-gram model.

The LDA topic model was trained in 1000 iterations with parameter alpha set to 50.0 and parameter beta to 0.01. The vectors corresponding to the words were normalized by dividing each value by the total frequency of the word.

A custom implementation was used to calculate HAL-like word vectors. We used a symmetric, flat window including 5 words on each side; then we applied a positive pointwise mutual information transformation to the resulting co-occurrence matrix; finally all words with frequency lower than 5 were removed from the corpus before training.

We trained a skip-gram model using a set of fairly standard settings: a window of 5 words and a starting learning rate of 0.025. The downsampling parameter was set to 1e-3, and hierarchical softmax was used when training the model. As in the case of the HAL model, all words with a frequency lower than 5 were discarded when training the model.

Only words that were simultaneously included in all three word vector models, in the rating sets, and in the norms for orthographic (Balota et al., 2007) and semantic density measures (Shaoul & Westbury, 2010) were used during the extrapolation procedure. This resulted in datasets containing 20,265 words for AoA, 20,994 for concreteness, and 12,531 words for affective ratings.

Extrapolation methods

To predict ratings using the k-nearest neighbours model, for each word in the test set we identified the 30 most similar (measured with cosine distance) words in the training set. This parameter (k = 30) was set to a value found by Bestgen and Vincze (2013) to be optimal in their extrapolations. The mean rating of these words was assigned to the target word as an extrapolated rating.

The random forest model was trained with 100 estimators. Taking advantage of the flexibility of this method with respect to number of predictors used, we also trained random forest models with additional predictors: $\log_{10}$ of word frequency and dominant part of speech. In the case of semantic vectors, the score obtained on each of the dimensions was used as a separate predictor.

RESULTS

General results

To measure the prediction accuracy of the different models, we first examined the correlations between the reconstructed and the original ratings (see Table 1), averaged across the 10 iterations.

Table 1. Correlations between the original ratings and the ratings extrapolated with different models trained on 25% of the full dataset (average of ten iterations).

			Variable				
Method	*Word vectors*	*Additional predictors*	*AoA*	*Conc*	*Arousal*	*Domin*	*Valence*
LM		wf	.621	.165	.054	.157	.174
		wf, len, old20	.635	.37	.143	.164	.178
		wf, len, old20, inc	.641	.371	.183	.195	.21
KNN	lsa		.540	.525	.299	.342	.412
	tm		.545	.647	.358	.370	.443
	hal		**.737**	.758	.44	.568	.661
	sg		.715	**.796**	**.478**	**.595**	**.694**
Random forest	lsa		.711	.609	.317	.395	.448
	tm		.695	.672	.374	.421	.500
	sg		.688	.723	.406	.543	.615
	lsa	wf	.730	.611	.315	.395	.454
	tm		.733	.681	.376	.422	.507
	sg		.730	.724	.407	.544	.618
	lsa	wf + pos	.731	.711	.318	.397	.453
	tm		.734	.746	.379	.422	.507
	sg		.730	.781	.406	.543	.616

Note*:* LM = linear model, KNN = *k*-nearest neighbours, wf = $\log_{10}$ of word frequency, lsa = Latent Semantic Analysis, tm = topic model, sg = skip-gram. Due to the large number of observations differences in correlations as small as .015 are statistically significant.

The baseline models with different combinations of predictors that did not include word vectors managed to predict the ratings to a limited extent compared to the models using semantic spaces. The ratings predicted by the model including only word frequencies correlated .621 with original AoA ratings but predictions of the simplest baseline model were much less successful for other variables and did not reach the level of .2 correlation for any other variable. Including information about orthographic properties of a word (length and neighbourhood density) more than doubled the correlation between the extrapolated and the original ratings for concreteness but affected accuracy of the extrapolation for the other variables to a much lesser extent. Adding a measure of semantic neighbourhood density increased the correlations most strongly for the affective ratings, but the accuracy of the extrapolation for these variables remained very low.

For all variables, we obtained higher correlations with the original ratings when the extrapolation methods took into account semantic information from the word vectors.

For all variables, the correlations obtained with the k-nearest neighbours outperformed those based on the random forest models. When the k-nearest neighbours method was used, HAL and skip-gram gave higher correlations than LSA and topic models. The highest correlation obtained was .737 for AoA (k-nearest neighbours with HAL word vectors), .796 for concreteness, .478 for arousal, .595 for dominance, and .694 for valence (k-nearest neighbours with the skip-gram word vectors).

In the case of AoA and concreteness, the ratings extrapolated with random forest were close to those extrapolated with the k-nearest

neighbours when word frequency and part of speech (POS) information were included as additional predictors. Extrapolation of AoA with random forest improved most when word frequency was added to the model based on word vectors only. For concreteness, including POS information increased the correlations most. For all affective ratings, including word frequency or POS among the random forest predictors had little effect on the accuracy of the predictions.

Usefulness of extrapolated AoA ratings

Variance explained in lexical decision task reaction times

Because the lexical decision task (LDT) is one of the most popular tasks in psycholinguistics, we looked at how much of the variance in reaction times (RTs) collected in the British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2011) and in the English Lexicon Project (ELP; Balota et al., 2007) is accounted for by reconstructed ratings in comparison to the variance explained by the original human ratings for these variables.

In our analysis we jointly entered extrapolated ratings from the test sets of all extrapolation iterations for words that were also included in the BLP and ELP. The resulting dataset included 10,471 unique words for AoA (about 7.5 extrapolations per word), 10,828 unique words for concreteness (about 7.26 extrapolations per word), and 7507 unique words for the affective variables (about 7.5 extrapolations per word). First, we created a baseline to which models including extrapolated ratings should be compared by fitting a model containing only $\log_{10}$ of word frequency as a predictor to the reaction times. Second, we created a model containing both $\log_{10}$ of word frequency and the original ratings as predictors.

Next, we fitted regression models including $\log_{10}$ of word frequency and the ratings predicted with different methods. The results of these analyses are shown in Table 2.

In general, we observed a consistent pattern for the different methods across ELP and BLP. However, the pattern of variance explained by the extrapolated ratings did not strictly follow the pattern of absolute correlations between the extrapolated and the original ratings. As could be expected, $\log_{10}$ of word frequency explained a large fraction (over 42%) of the total variance in RTs. When the original AoA ratings were included in the model, the percentage of variance accounted for increased by 3.21% for BLP and 3.37% for ELP. When we added the original concreteness ratings to the model, the percentage of explained variance increased by 0.38% for BLP and 0.35% for ELP. The effects of adding the affective variables were small and did not exceed 0.5% in any case.

Table 2. Percentage of variance explained by linear models with different sets of predictors.

			Additional variance explained [%]									
			AoA		*Conc*		*Arousal*		*Dominance*		*Valence*	
Method	*Word vectors*	*Additional predictors*	*BLP*	*ELP*	*BLP*	*ELP*	*BLP*	*ELP*	*BLP*	*ELP*	*BLP*	*ELP*
(baseline model)			**47.97**	**43.05**	**48.12**	**42.01**	**45.40**	**37.97**	**45.18**	**37.72**	**45.15**	**38.18**
(baseline + original ratings)			3.21	3.37	0.38	0.35	0.00	0.11	0.34	0.43	0.28	0.32
LM		wf	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		wf, len, old20	0.11	0.93	0.03	0.04	0.28	1.45	0.38	0.42	0.28	**1.13**
		wf, len, old20, inc	0.01	0.17	0.07	0.01	**0.89**	**2.18**	**1.02**	**1.39**	0.37	0.33
KNN	lsa		0.01	0.01	0.10	0.14	0.14	0.15	0.22	0.20	0.33	0.29
	tm		0.28	0.32	0.14	0.28	0.24	0.41	0.43	0.46	0.55	0.63
	hal		1.20	1.02	0.14	0.15	0.48	0.63	0.73	0.59	**0.67**	0.48
	sg		0.40	0.26	**0.25**	**0.31**	0.31	0.52	0.28	0.27	0.32	0.32
Random forest	lsa		1.05	0.96	0.00	0.07	0.06	0.10	0.70	0.48	0.65	0.42
	tm		**1.38**	**1.49**	0.01	0.07	0.10	0.23	0.65	0.68	0.48	0.48
	sg		0.39	0.26	0.07	0.12	0.16	0.31	0.14	0.13	0.19	0.21
	lsa	wf	0.74	0.67	0.00	0.07	0.04	0.10	0.68	0.45	0.59	0.43
	tm		1.13	1.20	0.01	0.07	0.08	0.20	0.60	0.68	0.36	0.36
	sg		0.81	0.59	0.07	0.13	0.14	0.31	0.14	0.15	0.19	0.22
	lsa	wf, pos	0.74	0.66	0.01	0.04	0.05	0.10	0.68	0.42	0.60	0.43
	tm		1.16	1.19	0.00	0.04	0.08	0.20	0.61	0.67	0.35	0.35
	sg		0.82	0.58	0.04	0.13	0.16	0.31	0.11	0.13	0.17	0.21

Note. The first row shows how much of the variance in reaction times taken from British Lexicon Project (BLP) and English Lexicon Project (ELP) is explained by a linear model with $\log_{10}$ of word frequency as the only predictor. The following rows show additional variance explained when original and extrapolated ratings for age of acquisition (AoA), concreteness (conc), and affective ratings were added to the model. Column 1 specifies the extrapolation method, column 2 shows the type of word vectors used with the method (lsa = latent semantic analysis; tm = topic model; hal = hyperspace analogue to language; sg = skip-gram), column 3 lists the additional predictors used when extrapolating the variable (wf = $\log_{10}$ of word frequency; len = word length, i.e., number of letters; old20 = orthographic Levenshtein distance 20; inc = inverse N count; pos = part of speech).

In the case of concreteness, the extrapolated ratings that had the highest correlation with original ratings were the ones that also explained most of the variance in RTs (0.25% above the baseline model for BLP and 0.31% for ELP). For AoA a different pattern emerged. For this variable the ratings extrapolated with a random forest combined with topic models without including any additional predictors gave the largest improvement compared to the baseline model (1.38% for BLP and 1.49% for ELP). Interestingly, this was not the extrapolation method that correlated most strongly with the original ratings, and, although the ratings extrapolated with the k-nearest neighbours combined with HAL-like word vectors also predicted a large fraction of the variance (1.20% for BLP and 1.02% for ELP), in general the pattern of explained variance in RTs did not strictly follow the pattern observed in absolute correlations with the original ratings. For example, although ratings extrapolated with skip-gram word vectors and k-nearest neighbours correlated more strongly with the original ratings than those based on random forest and topic models, the former explained 3.5 times less variance in RTs for BLP and 5.7 times less for ELP than the latter.

Surprisingly, for the affective ratings we found that many of the extrapolated variables explained more additional variance in the RTs than the original ratings when added to the linear model including word frequencies. Moreover, predicted ratings that had some of the weakest correlations with the original ratings seemed to explain the largest fraction of the variance in the RTs. For arousal, the linear model including information about word frequency, length, and orthographic and semantic neighbourhood density predicted ratings that correlated only 0.183 with the original ratings but, when these

ratings were used to predict RTs, they improved explained variance by 0.89% for BLP and 2.18% for ELP, while the original arousal ratings made hardly any difference in the explained variance. For dominance, the same model gave predictions that correlated .195 with the original ratings but improved the explained variance by 1.02% for BLP and 1.39% for ELP, while the original ratings explained only 0.34% additional variance for BLP and 0.43% for ELP. In the case of valence, variance in RTs taken from BLP was best accounted for by the ratings extrapolated with the k-nearest neighbours and word vectors obtained with HAL-like method (0.67% extra explained variance), and variance in RTs taken from ELP was best explained by the ratings extrapolated with the linear model including only word frequency, word length, and OLD20 as predictors (1.13% of additional explained variance). At the same time, the original ratings for valence explained only 0.28% extra variance for BLP and 0.32% for ELP. Improvements of the explained variance above the level explained by the original affective ratings were strongest in the case of the simple linear models but the k-nearest neighbours and random forest methods also produced ratings that explained more variance in lexical decision RTs than the original ratings.

In order to explain the surprising effects regarding explained variance in lexical decision RTs, we conducted an additional analysis in which we investigated whether the extrapolation procedures could introduce artefacts to the data that could easily be identified with effects of some of the well-known psycholinguistic variables. In order to do that, we looked at the correlation structure of the original and reconstructed ratings with variables known to influence performance in psycholinguistic tasks: length, OLD20, word frequency, semantic

neighbourhood density (inverse N count), and ratings for all the variables that we were extrapolating in the current study. We reasoned that, in order to represent the same theoretical construct, the extrapolated ratings should not only correlate with the original variables as strongly as possible but also have similar correlations with other variables as the original ratings. When looking at effects as small as the effects of affective variables on lexical decision RTs, even small artefacts could distort the conclusion that would be reached based on a particular analysis. We indeed observed that the extrapolated ratings had a different correlation structure than the original ratings.

As could be expected based on the patterns of explained variance in lexical decision RTs, the most striking discrepancies in correlation structure were observed for the affective variables. For arousal, the extrapolated ratings that explained the largest fraction of the variance in RTs (linear model with $\log_{10}$ of word frequency, length, and orthographic and semantic neighbourhood density measures) correlated .5 with OLD20 and .51 with word length. These correlations were much higher than the correlation of .1 for both OLD20 and length in the case of original ratings. We observed a similar pattern when we looked at the dominance ratings extrapolated with this method. In this case, although the differences in correlations were smaller: −.18 for length (−.04 for the original ratings) and −.34 for OLD20 (−.07 for the original ratings), the differences for correlations with word frequencies (.78 for the extrapolated ratings and .16 for the original ratings) and inverse N count (−.91 for the extrapolated ratings and −.18 for the original ratings) were very high. In the case of valence, the ratings extrapolated with a model that

explained the largest fraction of the variance in RTs from ELP (linear model with $\log_{10}$ of word frequency, length, and OLD20 as predictors) had much higher correlations than the original ratings with $\log_{10}$ of word frequencies (.97, .17 in the original ratings), inverse N count (−.58, −.20 in the original ratings), and AoA ratings (−.50, −.22 in the original ratings). Although such discrepancies were strongest for ratings extrapolated with the linear models, we observed similar tendencies in the ratings extrapolated using semantic vectors. For instance, for the valence ratings extrapolated using k-nearest neighbours and HAL-like word vectors, the correlation with length was −.15 (−.02 for the original ratings), with OLD20 −.16 (−.03 for the original ratings), with word frequency .35 (.17 for the original ratings), with inverse N count −.32 (−.20 for the original ratings), and with AoA −.33 (−.22 for the original ratings).

Although these results suggest that some artefacts are present in the extrapolated ratings, it is possible that there are further confounds that can not be easily identified with one of the variables that we considered in our analysis of the correlation structure. Because of that, we decided to conduct one more analysis: We decorrelated the extrapolated ratings with the original ratings by fitting linear models in which we predicted the extrapolated ratings based on the original ratings and considered residuals of such a model as a representation of what the ratings capture in addition to the variance that they share with the original ratings. Next, we checked whether the residuals of the extrapolated ratings can still predict a meaningful amount of variance in behavioural data when they are added to a linear model in which we entered BLP RTs as a dependent variable and word frequency as an independent variable. If that would be the case, it

would indicate that variance that is present in the extrapolated ratings but that cannot be identified with the original ratings can be predictive of behavioural variables. In such a case, if the extrapolated ratings would be used in a hypothetical analysis, we could reach conclusions other than we would reach based on the original ratings because of such a confound.

For all variables we found that the residuals of the extrapolated ratings still explain a meaningful amount of variance above what can be explained by word frequencies alone. This was the case not only for the ratings extrapolated using the linear models but also for some of the ratings that were extrapolated using semantic spaces. For instance, the residuals of the ratings extrapolated with the k-nearest neighbours method and HAL-like word vectors explained 0.21% additional variance in RTs in the case of AoA, 0.56% for arousal, 0.32% for dominance, and 0.30% for valence, and the ratings extrapolated with random forest and topic model word vectors explained 0.16% extra variance in the case of concreteness, 0.10% in the case of arousal, 0.36% in the case of dominance, and 0.24% in the case of valence.

Categorizetion of the extrapolated variables

In psycholinguistic research, variables that can be measured on a continuous scale are often dichotomized or binned. Therefore we compared how binning based on extrapolated AoA ratings compared to binning using the original ratings. To conduct this analysis we again used the full set of words extrapolated in all 10 iterations. In order to obtain a benchmark for the performance of the extrapolation

procedures, we used two random splits of the data collected by Kuperman et al. (2012).[4]

Applying a dichotomization or binning procedure to the ratings is equivalent to reformulating the evaluation from a regression problem where the variables are considered on a continuous scale to a classification problem where the outcomes take discrete values. We decided to test the quality of the classification based on extrapolated ratings by using two procedures:

1. Dichotomization of the set of words by splitting it at different points across the entire range of the original ratings. This is equivalent to asking how precise our predictions would be if we used extrapolated ratings to predict which words were already acquired before a certain age. In order to answer this question, we split the full dataset in bins corresponding to each year of life (from 1 to 24). All words with an original AoA rating below that age were considered as positive cases (already-acquired words) and the remaining words as negative cases (words that were not yet acquired). All the words that should have been acquired at that age according to the extrapolated ratings were considered to be classified as already-acquired words, and all remaining words as words that were not yet acquired.
2. Splitting the full dataset into bins corresponding to deciles of AoA, which is equivalent to asking how precisely we can predict that a given set of words will be the next 10% of words acquired

[4] We used a dataset obtained from the authors of the original study (Kuperman et al., 2012). The dataset did not correspond perfectly to the one on which the published ratings were based and which was used to train the models but had a very high correlation ($r = .96$) with that dataset. A total of 705 words that were not included in the dataset were excluded from the analysis.

after a given percentage of words was already acquired. For example, evaluating how precisely we can predict words in the third decile corresponds to the precision of making a prediction about a set of words that will be acquired after 20% of all words were already acquired but before the remaining 70% of words. In order to conduct this analysis, we binned the words based on the deciles in the original set of AoA ratings and, separately, in the extrapolated ratings. Next, we evaluated the classification performance for each of the bins. All words acquired in that bin according to the original ratings were considered as positive cases, and all remaining words as negative cases. All words included in a corresponding bin of the extrapolated ratings were considered to be positive cases, and the remaining words were considered to be negative cases.

The two evaluation procedures can be seen as binary classification problems. The overall result of the classification can be represented in a 2 × 2 matrix, which includes: true positives (correctly classified positive cases; TP), true negatives (correctly classified negative cases; TN), false positives (negative cases incorrectly labelled as positive; FP), and false negatives (positive cases incorrectly labelled as negative; FN). Based on these classification results, we calculated a set of metrics that are commonly used to measure performance of classification methods:

Accuracy

Accuracy represents the fraction of correctly classified positive and negative examples.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Note that this metric is insufficient if there is a difference in the size of TP and TN classes. For example, if only 5% of the cases in the original dataset would be the positive cases, a method that labels all cases as negative, irrespective of the input, would achieve 95% accuracy. To correct for this possibility, we calculated a set of additional metrics.

Sensitivity and precision

Precision represents the fraction of cases that were classified as positive and were also positive in the original dataset.

$$precision = \frac{TP}{TP + FP}$$

Sensitivity represents the fraction of all positive cases in the original dataset that were correctly classified as positive.

$$sensitivity = \frac{TP}{TP + TN}$$

F1-score

$$F_1 = 2 * \left(\frac{precision * sensitivity}{precision + sensitivity}\right)$$

F1-score (Rijsbergen, 1979) is a harmonic mean of precision and sensitivity. It simultaneously takes into account both how many of the relevant cases were correctly identified by the method and how many nonrelevant cases were mistakenly labelled as positive.

Figure 1 shows the metrics calculated for the first classification

procedure in which the dataset was split in two groups at different points of the range of the original ratings.

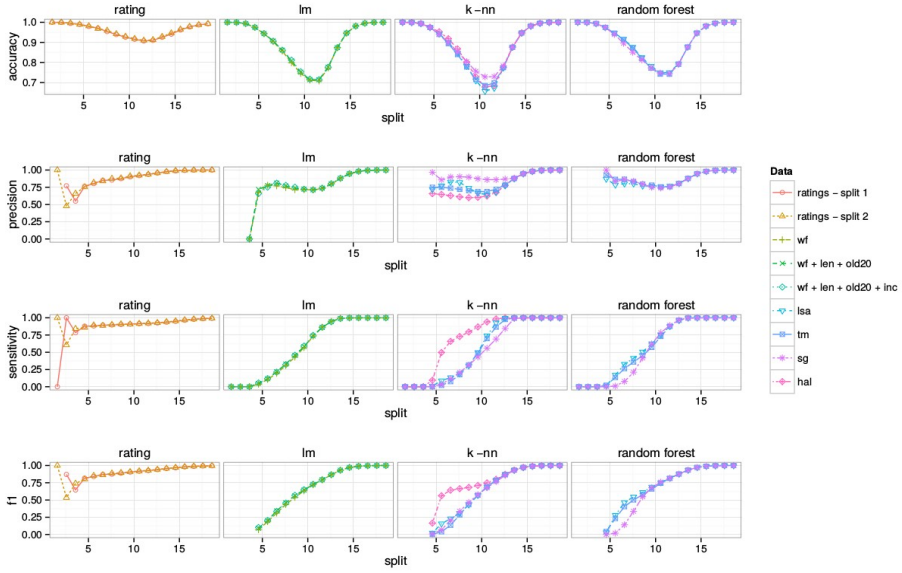

Figure 1. Performance metrics representing the quality of the classification when splits into two groups were made at different age of acquisition values. Each row represents a different performance metric. The leftmost column shows the metrics calculated for the ratings based on two splits of the human ratings dataset. The remaining columns show the classification performance metrics for the different extrapolation methods. The different lines in the figure represent different sets of predictors that were used to make the extrapolation. The extrapolations in which the random forest method was used with additional predictors were removed from the plot because they followed very similar patterns to the extrapolations shown. lm = linear model; k-nn = k-nearest neighbours; lsa = latent semantic analysis; tm = topic model; hal = hyperspace analogue to language; sg = skip-gram; wf = $\log_{10}$ of word frequency; len = word length; old20 = orthographic Levenshtein distance 20; inc = inverse N count.

As can be seen on the figure, the closer to the boundaries of the range, the higher the accuracy. This probably reflects the fact that,

when all words are taken into account, it is easier to make accurate predictions close to the boundaries of a scale. As such, it can be considered an artefact of different prior probabilities for different classes. For precision we can observe that ratings extrapolated with k-nearest neighbours and with the random forest method stay at a rather high level for most half-splits across the entire range of the AoA ratings. At the same time, sensitivity starts at a very low level and rapidly increases until the age of 15. This pattern of sensitivity and precision metrics probably reflects the distortion of the scale that happens when the extrapolation procedures are applied. For example, when applying the k-nearest neighbours method, on average words are shifted towards the mean age. As a result, the extrapolation method has a tendency to overestimate AoA for early-acquired words. Because of that, the precision is high: Few words that are not yet acquired according to the original ratings are classified as already acquired. At the same time, the method fails to identify words that are acquired at an early age according to the original ratings. The F1-score shows the overall performance of the extrapolation methods with different splits. Because it involves a product of precision and sensitivity, this metric stays at a low level due to low sensitivity despite high precision. This pattern can be contrasted with the high precision and sensitivity across the entire range for the two sets of ratings calculated based on half-splits of the full ratings dataset. This result shows that the usefulness of extrapolated ratings may be limited when accurate identification of early-acquired as opposed to late-acquired words is necessary unless the split is made at a relatively high age.

As shown in Figure 2, splitting the dataset by AoA decile produced a much more regular pattern across all the metrics. Because

binning into different deciles depends on ranks of words and not on the absolute AoA values assigned to different words, this classification procedure is not affected by the distorted scale. All metrics show that the quality of binning is better for the extreme deciles. Most probably, this is caused by the fact that the extreme deciles contain all the words with potentially unbounded range at one of the sides, which increases the accuracy by allowing methods to assign a word to the correct bin even if the prediction is inaccurate in terms of an absolute value. All metrics stayed at a rather low value for most of the nonextreme deciles. This result shows that the extrapolation methods may not be accurate enough to be used for assigning words to classes spanning a limited range.

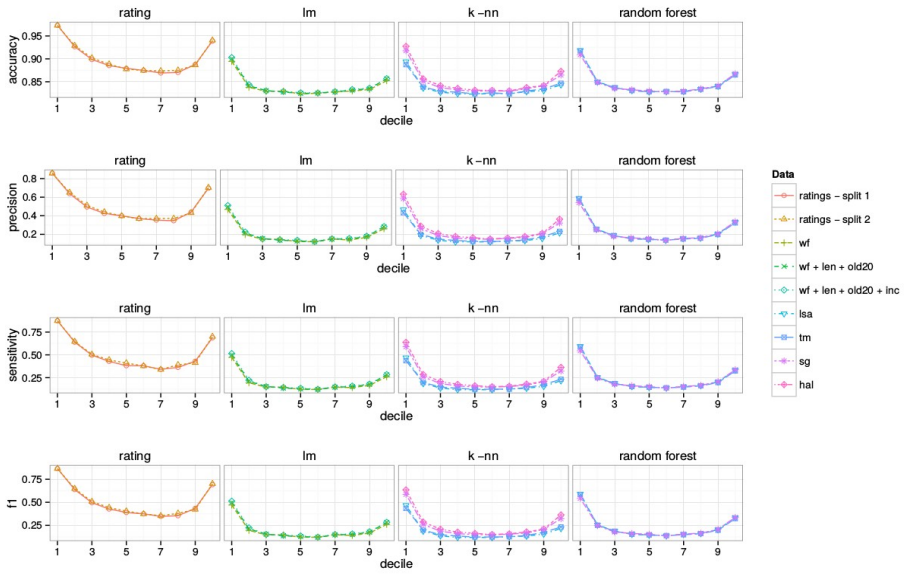

Figure 2. Values for different performance metrics representing quality of the classification into individual bins. The different lines in the figure represent different sets of predictors that were used to make the extrapolation. lm = linear model; k-nn = k-nearest neighbours; lsa = latent semantic analysis; tm = topic model; hal = hyperspace analogue to language; sg = skip-gram; wf = log$_{10}$ of

word frequency; len = word length; old20 = orthographic Levenshtein distance 20; inc = inverse N count.

Trainingset size and prediction accuracy

In addition to the analyses reported so far, we investigated how prediction accuracy depends on the size of the training set. We ran 10 iterations of the extrapolation procedures, with splits of 10%, 25%, 50%, 75%, and 90% of the data in the training set and, respectively, the remaining 90%, 75%, 50%, 25%, and 10% in the test set.

The results of this analysis are shown in Figure 3. In general, we observed a steady increase in the accuracy of predicted ratings up to a training set size of 10,000 in the case of the methods that made use of the semantic vectors. As could be expected, the larger the training set the smaller further increases in the accuracy of the predictions.

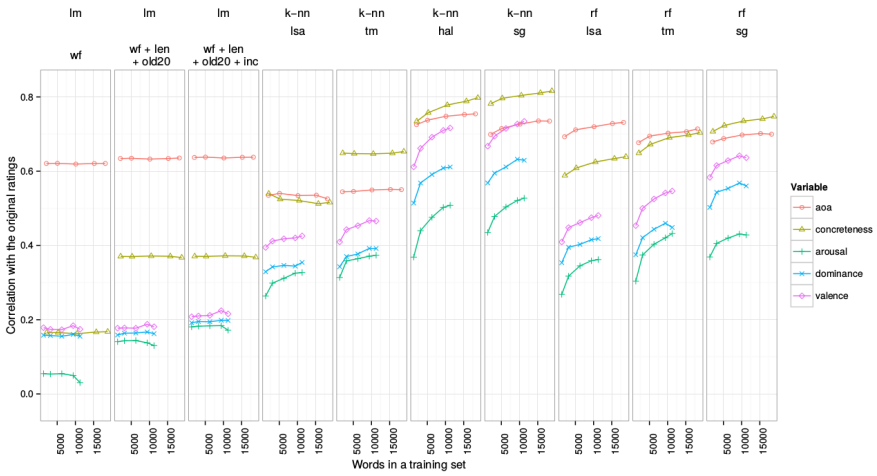

Figure 3. Correlations between original ratings and ratings extrapolated on the basis of different numbers of words included in the training set (average of 10 iterations). The different lines in the figure represent different extrapolated variables. lm = linear model; k-nn = k-nearest neighbours; rf = random forest;

lsa = latent semantic analysis; tm = topic model; hal = hyperspace analogue to language; sg = skip-gram; wf = log_{10} of word frequency; len = word length; old20 = orthographic Levenshtein distance 20; inc = inverse N count; aoa = age of acquisition.

DISCUSSION

We conducted a systematic comparison of two extrapolation methods using different vector representations of words to predict ratings of psycholinguistic variables.

Our analyses showed that the k-nearest neighbours used with word vectors from the skip-gram and HAL-like model give the most accurate predictions. This is true especially for variables where the semantic component plays a primary role. On the other hand, when other predictors can bring important information to the model, the random forest method is the most convenient to use. Because both k-nearest neighbours and random forests have their own strengths, it would be interesting to find a way to create a hybrid technique that is able to make use of the strengths of each of the methods.

At the same time, we have shown that the usefulness of ratings extrapolated with currently available methods may be limited. In particular, the result of our analysis in which we predicted lexical decision RTs using the extrapolated ratings gave some surprising results. It also seems problematic to rely on extrapolated ratings when dichotomizing or binning words. Although we conducted the analysis in which we categorized otherwise continuous data only for AoA it may be expected that the result would be even worse for other variables such as affective ratings, since AoA was the variable for

which the extrapolation methods produced relatively high correlations with the original ratings.

Our analyses clearly show that reporting the correlation between the original and the extrapolated variables is not sufficient to evaluate their usefulness. Even if extrapolated ratings share a large fraction of the variance with the original ratings there is still a part of the variance that does not reflect the original ratings, and we cannot assume that this variance is just random, unsystematic noise. In contrast to the half-splits of human data, in which case we can safely assume that in both splits the uncorrelated part of the variance have similar statistical structure, we cannot make such an assumption in the case of comparing the product of statistical models (extrapolated ratings) with human ratings.

It is easy to understand how the artefacts can arise in the case of extrapolations based on linear models. Due to the nature of this method, the predictions are always proportional to the values of the predictors. As a result, the predictors can “leak” into the extrapolated variables.

For instance, let us consider a hypothetical case where we would train a linear model that would predict ratings as a combination of word frequency and OLD20 with respective coefficients of .5 and . 4. In this case, if we extrapolated ratings for two words that have equal frequency, the word with higher OLD20 would always obtain a higher rating. Because the predictions are usually imperfect, there is always some error in the predictions, and, because the variance that does not reflect the original ratings is not just random noise, but rather is strongly correlated with OLD20, the error would be also correlated with OLD20. Although it is more difficult to explain how such effects

can arise in the case of the k-nearest neighbours methods and the random forest methods, it has already been demonstrated that some properties of the semantic space may be associated with well-known psycholinguistic variables. For example, it has been shown that some of the semantic neighbourhood density measures can strongly correlate with word frequencies even if the frequencies are not explicitly encoded in the semantic space (Shaoul & Westbury, 2006). Similarly, implicit properties of the semantic spaces can lead to introducing artefacts to the extrapolated ratings.

Of course, the higher the correlation of extrapolated ratings with the original ratings, the less room for artefacts; we indeed observed that the artefacts were generally smaller in the case of extrapolated ratings that correlated more strongly with the original ratings. At the same time, it seems important that in the case of extrapolated ratings we are not looking at the original phenomenon but rather at the output of a statistical model. In such case it may be impossible to disentangle patterns in the data that arise due to properties of the phenomenon from those that arise due to properties of the model itself. This aspect of the extrapolated ratings can make it problematic to use them interchangeably with the human ratings or draw strong conclusions based on such ratings.

Despite these limitations, the extrapolated variables still seem to have some important applications. For instance, the extent to which different extrapolation methods with different predictors are successful in predicting ratings can potentially inform us about the psycholinguistic variables. For instance, the fact that co-occurrence similarity between words explains a nontrivial part of variance in AoA ratings could suggest that semantically related words are acquired

around the same age. The same logic can be applied to the other variables, although, as was already reported by Bestgen and Vincze (2012), co-occurrence models often model antonyms as close neighbours in a vector space. It would be interesting to look at how this problem can affect extrapolation of different variables. For example, love and hate are obviously on the opposite sides of the valence continuum, so modelling them as close semantic neighbours may be a problem for extrapolating valence, but this problem should affect to a smaller extent variables such as AoA or concreteness, as there is no reason why there would be a strong tendency to acquire antonymous words at very different age or why one of the words in the antonym pair would be more concrete than the other.

In addition, the accuracy of extrapolation procedures using different word vector representations can be informative about the word vector representations themselves. Although we used models based on statistical distributions of words in a language as approximate representations of semantics, different models may capture its different aspects. For instance, apparently in our study the word vectors based on narrow windows (HAL-like model and skip-gram model) performed better than the bag-of-words models and perhaps such vectors allow us to model semantic similarity in a way that better corresponds to that reflected in psycholinguistic variables. It also seems plausible that the high correlations obtained using the skip-gram model can be simply explained by it being better at estimating word similarities (Baroni et al., 2014).

We have shown that increasing the size of the training set gives diminishing improvements to prediction accuracy as the training set gets larger. This means that, at least to some extent, extrapolation of

variables can be already applied even if the sets of seed ratings currently available are relatively small. On the other hand, together with rather disappointing results of the evaluation of the practical usefulness of extrapolated variables, it shows that further developments are necessary to allow for radically improved accuracy of the extrapolation procedures.

Because in the current study we used large sets of ratings, our results should generalize well across the entire lexicon. Despite that, the fact that the extrapolation methods as well as word vector representations require parameters to be specified during training may hamper the generalizability of our conclusions. Because the methods are computationally demanding, it seems implausible to try to cover the entire parameter space of all the methods. At the same time, there is no guarantee that what is found with one parameter setting would be true for another parameter setting. Especially there is no guarantee that we did not choose a more optimal set of parameters for one method than for the other methods. The result also depends on the corpus that was used to train the models and the way in which it was preprocessed. There is a possibility that the subtitle corpus we used may be suboptimal for the purposes of distributional semantics, which may have reduced the performance of the extrapolation methods. Indeed, some of the correlations reported in the literature (e.g., Recchia & Louwerse, 2014) were higher than the ones we found. However, it is difficult to make direct comparisons across studies as the sets of ratings, their sizes, proportions of the training and test sets, and approaches to cross-validation vary across studies. Moreover, the differences in the correlations reported across studies are not large enough to expect that using a different corpus would lead to

qualitatively different conclusions from the ones we reached here. Also there is no reason to believe that the overall pattern of relative efficacy between the different methods of extrapolation and the techniques of constructing word vectors would be different. Nevertheless, it would be interesting to look at the corpus effects in future studies of this type.

An interesting problem to address in future research is how we can optimize our data collection process to collect ratings, so that they become maximally informative for the extrapolation methods. If an optimal set of seed words would increase the accuracy of the extrapolation methods, it would be good to know this.

Finally, given recent developments in computational linguistics, it would be interesting to explore the possibilities of cross-language extrapolation of psycholinguistic variables. It was recently shown that it is possible to learn a linear mapping between vector spaces of two languages (Mikolov, Le, & Sutskever, 2013). This means that, in addition to word properties in a given language, we could use information from other languages when extrapolating ratings (e.g., use sets of ratings that were already collected for English to predict ratings for other languages).

SUPPLEMENTAL MATERIAL

Supplemental content (part of speech tags) is available via the "Supplemental" tab on the article's online page (http://dx.doi.org/10.1080/17470218.2014.988735).

REFERENCES

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). *Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors*. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 1). Retrieved from http://clic.cimec.unitn.it/marco/publications/acl2014/baroni-etal-countpredict-acl2014.pdf

Bestgen, Y. (2002). Détermination de la valence affective de termes dans de grands corpus de textes [Determination of the emotional valence of terms in large corpora]. In Y. Toussaint, & C. Nedellec (Eds.), *Actes du Colloque International sur la Fouille de Texte CIFT '02* (pp. 81-94). Nancy, France: INRIA.

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, *44*(4), 998–1006. doi:10.3758/s13428-012-0195-z

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Bradley, M., & Lang, P. (1999). *Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings (Technical Report No. C-1)*. Gainesville, FL: University of Florida, NIMH Center for Research in Psychophysiology.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Broder, A. Z. (1997). *On the resemblance and containment of documents*. Proceedings of Compression and Complexity of Sequences 1997 (pp. 21–29). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=666900

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition, 13*(7–8), 992–1011. doi:10.1080/13506280544000165

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. doi:10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911. doi:10.3758/s13428-013-0403-5

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. doi:10.1080/14640748108400805

Feng, S., Cai, Z., Crossley, S., & McNamara, D. S. (2011). *Simulating Human Ratings on Word Concreteness*. *Twenty-Fourth International FLAIRS Conference*. Retrieved from http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS11/paper/viewPDFInterstitial/2644/3035

Fix, E., & Hodges, J. (1951). Discriminatory analysis, nonparametric discrimination: Consistency properties. *US Air Force School of Aviation Medicine*, *Technical Report 4*(3).

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.*, *6*(6), 721–741. doi:10.1109/TPAMI.1984.4767596

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, *12*(4), 395–427. doi:10.3758/BF03201693

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*(2), 183–233.

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. doi:10.3758/BRM.42.3.643

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2011). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. doi:10.3758/s13428-011-0118-4

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. doi:10.3758/s13428-012-0210-4

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review; Psychological Review*, *104*(2), 211–240.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2), 203–208.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. Retrieved from http://arxiv.org/abs/1301.3781

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. *arXiv:1309.4168 [cs]*. Retrieved from http://arxiv.org/abs/1309.4168

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, *3*, 235–244.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods, 41*(3), 647–656. doi:10.3758/BRM.41.3.647

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology*, *68*(5), 1584–1598. doi:10.1080/17470218.2014.941296

Rijsbergen, C. J. V. (1979). *Information retrieval*. Butterworths.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. doi:10.1038/323533a0

Sahlgren, M. (2006). *The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. (Doctoral dissertation, Stockholm University). Retrieved from http://eprints.sics.se/437/1/TheWordSpaceModel.pdf.

Shaoul, C., & Westbury, C. (2006). Word frequency effects in high-dimensional co-occurrence models: A new approach. *Behavior Research Methods*, *38*(2), 190–195. doi:10.3758/BF03192768

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods, 42*(2), 393–413. doi:10.3758/BRM.42.2.393

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. S. McNamara, S. Dennis & W. Kintch (Eds.), *Handbook of Latent Semantic Analysis* (pp. 424–440). Hillsdale, NJ: Erlbaum.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). *Feature-rich part-of-speech tagging with a cyclic dependency network.* Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 173–180). Association for Computational Linguistics.

Toutanova, K., & Manning, C. D. (2000). *Enriching the knowledge sources used in a maximum entropy part-of-speech tagger*. Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13 (pp. 63–70). Association for Computational Linguistics.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191-1207. doi:10.3758/s13428-012-0314-x

Westbury, C. (2013). You Can't Drink a Word: Lexical and Individual Emotionality Affect Subjective Familiarity Judgments. *Journal of Psycholinguistic Research, 43*(5), 631-49. doi:10.1007/s10936-013-9266-2

Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it, now you don't: on emotion, context, and the algorithmic prediction of human imageability judgments. *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00991

Chapter 7. General discussion

The goal of this dissertation was to make use of the increased availability of digital materials to develop new resources for use in psycholinguistic research, to improve the methodology of creating such resources, and to exploit these resources to advance psycholinguistic theory.

The first empirical chapter of this dissertation presented a new set of word frequency norms for British English and demonstrated that the rapidly increasing availability of textual materials makes it possible to compile specialized text corpora that better approximate variants of language used by sub-populations of language speakers. Word frequency norms focused on British and American English were found to better predict human behavior recorded in datasets collected in the corresponding countries.

In chapter 3, I presented a set of new frequency norms for Polish. I proposed a more efficient procedure of evaluating frequency norms than the one based on megastudy data but found that the results of the evaluation may differ depending on the stimulus selection procedure. I also used a method of evaluating the frequency norms using web-based experimental data collection.

In chapter 4, I described an experiment in which data was collected from a demographically diverse group of participants and I showed how such datasets can be used to explain patterns of individual variability in the word frequency effect. In that chapter, I combined well-known properties of the word frequency distribution with basic principles of human learning to explain the observed patterns of changes associated with increased language exposure.

Moreover, I demonstrated that web-based experiments can be used to collect chronometric data from demographic groups that are difficult to recruit for participation in laboratory-based experiments.

In the last two empirical chapters of the thesis I extended the approach of combining megastudies with text-based measures to the semantic domain. In chapter 5, I discussed the different types of distributional semantics models as well as the relationships between these models and theories of learning such as the Rescorla-Wagner model (Rescorla & Wagner, 1972; Baayen, Milin, Filipovic Durdevic, Hendrix, & Marelli, 2011). Next, I demonstrated that they can indeed predict human behavior in psycholinguistically relevant tasks. As an important part of the analyses, I used a large dataset of semantic priming, extending the rationale used to evaluate the quality of word frequency norms to the evaluation of models of semantic relatedness. Contrary to earlier claims (Hutchison, Balota, Cortese, & Watson, 2008), I found that distributional semantics models can predict a significant percentage of the variance in semantic priming response times.

In chapter 6, I extended and carefully validated an approach to estimating human ratings based on semantic relatedness that was previously proposed in the literature (e.g., Bestgen & Vincze, 2012). Although the investigated extrapolation methods achieved high-correlations with human ratings, I found that this may be insufficient to use them as a replacement for the original ratings.

Overall, the studies included in my dissertation revealed that even for well-studied phenomena, such as the word frequency effect, the underlying processes are often more complex than what is typically revealed by small-scale studies. By using large datasets with

an increased number of stimuli and significant demographic variability, we can obtain a much more detailed view of psycholinguistic phenomena. As a first example, compiling a more specialized corpus of word frequencies confirmed that available word frequency norms are not equally predictive of behavioral data in all variants of English. Secondly, by simultaneously considering different corpus sizes and different populations of participants, I demonstrated that, if we consider a full spectrum of word frequencies, there is no definite answer to the question what the *minimal* size of a corpus should be to derive "good" word frequencies for psycholinguistic research. One reason is that increasing the size of the corpus invariably benefits how well word frequencies will predict behavioral data in the low frequency range, even though the size of the corpus is much less important for the high frequency words. Moreover, the part of the frequency range in which the word frequency effect is situated seems to shift with increased exposure to language. The situation is also complicated in the case of distributional semantics models because using different models trained on different text corpora works best for predicting performance in different tasks. Finally, I brought important nuances to using predictive methods to estimate human ratings. I showed that these methods produce ratings that strongly correlate with the ones collected from human participants but that they introduce statistical artifacts to the data and thus may not be a viable substitute for the human ratings.

IMPLICATIONS FOR THE USE OF TEXT CORPORA IN PSYCHOLINGUISTICS

The results presented in this dissertation show that both qualitative and quantitative aspects of text corpora are important for psycholinguistic research. Luckily, it is now easier than ever to compile corpora that are sufficiently large to be used for research and that, at the same time, more accurately reflect linguistic experience of language speakers.

The results of chapter 4 of the dissertation demonstrated that the size of the corpus matters much more for low- than for high-frequency words. In combination with the influence that increased exposure to language has on the word frequency effect, this implies that the size of the corpus is increasingly important for modeling the word frequency effect as participants become more linguistically experienced. This finding also offers a new perspective on somewhat puzzling effects reported in the third chapter. In our analysis of the Polish data we observed that word frequencies based on a corpus consisting of predominantly written materials predicted more of the variance in reaction times in the low-frequency range than did word frequencies based on corpus of film subtitles. If we consider that the subtitle corpus was the smaller one in this comparison, the fact that the size of the corpus is more important for the low frequency words naturally explains this result. Moreover, it also explains why we did not observe the same pattern when we conducted a similar analysis using the British Lexicon Project megastudy data (Keuleers, Lacey, Rastle, & Brysbaert, 2011). In this case the subtitle corpus used to derive measures was larger than the written text corpus (BNC), so

frequencies derived from this corpus would have stronger correlations with behavioral data for both the high- and the low-frequency words. This shows that, in addition to issues of corpus register, corpus size may be an important methodological consideration in psycholinguistic research, even when dealing with corpora of one hundred million words or more.

The results of the analyses conducted in chapter 4 show that it may be dangerous to follow the logic of factorial experiments and try to categorize words into high- and low- frequency groups. At least in terms of the reaction times in the lexical decision task, the boundary between these two categories would have to shift with increased linguistic experience of a participant. In other words, the word frequency effects shifts towards the lower-end of the frequency spectrum with increased experience. What is a high- and low-frequency word is impossible to define in absolute terms, but depends on who are the participants in the experiment. At the same time, the observation that for participants typically taking part in psychological experiments, the word frequency effect is situated in the much lower part of the continuum than was often assumed, is still valid (chapter 2).

On the other hand, the qualitative aspects of the text corpora are also important. It does not matter how precise the frequency estimates are if they are based on a sample from a non-representative frequency distribution. With respect to such qualitative aspects of text corpora, our results confirm the advantage of the frequency norms based on movie subtitles as compared to those based on other materials, providing that the subtitle corpus is sufficiently large. Moreover, in chapter 3 we have shown that it may be worthwhile to

try to approximate the variant of the language that is used by the participants more closely. We demonstrated that point for British and American English, but substantial within-language variability is present in many languages, so this result has implications beyond research conducted in English.

Keeping this in mind, it may be an interesting question for future research to investigate whether the variant of a language that the person experiences on a daily basis can be approximated even better. This would allow us to answer the question how much linguistic experience varies across different individuals and to what extent is it reflected in how we process language. This question may be especially interesting from a point of view of what we know about the statistics of language. It is known that content words are not distributed uniformly in text corpora but rather occur in bursts (Katz, 1996). Similarly, human linguistic experience is specialized and what is a low-frequency word for one person may actually be very frequent in the experience of another person (as it is likely the case with the word *corpus* for the readers of this dissertation). Given that we consume an increasing amount of text on-line and that Internet companies already track our Web-browsing habits to carefully analyze the content of the Web-pages to optimize their advertising campaigns, performing such analyses could become possible if researchers were able to gain access to such datasets or collect them on their own.

Outside of the domain of word frequencies, we looked at the effects associated with using different text corpora in the context of distributional semantics. Interestingly, we found that the tasks used to evaluate the models seem to determine which corpus is the most adequate for training. In the case of distributional semantics the

qualitative aspects of the data also tend to be important enough to outweigh even large differences in sizes of the corpora used to train the models. Importantly, in semantic priming the models trained on subtitle corpora predicted behavioral data better than much larger corpora not based on subtitles, showing that the subtitle corpora may also be useful in domains other than estimating word frequencies.

Implications for data acquisition methods

One of the goals of this thesis was to extend the megastudy approach by proposing new methods of collecting behavioral data. We approached this problem from two different angles. We investigated how we can make the data collection easier and less time and resource consuming by (a) conducting experiments in a web browser (chapter 3 and 4) running on a wide range of devices, (b) collecting the data in a smarter way by choosing the optimal set of stimuli to evaluate frequency norms (chapter 3), or (c) collecting ratings only for a small number of words and use information from distributional semantics models to estimate the ratings for the remaining words in a language (chapter 6).

Collecting more data using Web-based experiments generally proved to be good and more robust compared to the optimized data collection (chapter 3) or extrapolation (chapter 6) methods discussed in this thesis. The strategy of selecting the maximally differentiating sets of stimuli between text corpora to facilitate their evaluation indeed produced large differences in the performance of the compared frequency norms, but our study also made it very clear that the results of the evaluation may differ dramatically depending on which sets of stimuli are used. Therefore, it is safer to use more neutral methods of

selecting stimuli for such studies. In the fifth chapter of the dissertation the estimated ratings proved to have high correlations with the original ratings. However, the structure of the noise differed relative to the original ratings, which is problematic for using such ratings in psycholinguistic research. Still, the method may be useful for other applications, for instance in more engineering oriented tasks, such as sentiment analysis.

On-line data collection techniques worked very well for our purposes, both on a small (chapter 2) and a very large (chapter 4) scale. Even if the chronometric measurements collected in on-line studies are more noisy than those from carefully controlled laboratory-based experiments, there are two reasons to be optimistic about the possibility of obtaining reliable data. The first is a set of findings that show that the variability in human responses generally outweighs the variability because of timing inaccuracies (Damian, 2010), the precision of measurements in Web-browsers is high (Reimers & Stewart, 2014) and that many paradigms can be replicated in web browsers, even in relatively uncontrolled conditions (Crump, McDonnell, & Gureckis, 2013). In our case, variance in response times was also compensated by a dramatically larger number of observations, which resulted in reliable aggregated reaction times. It has been suggested that it may be useful to conduct coordinated experiments for a large number of languages and that the Web-based methodology may be ideal for such purposes (Myers, submitted). The two largely compatible datasets of vocabulary knowledge for English and Dutch, that we collected, may be a good start of such a lexical decision meta-megastudy. Two more datasets, compatible with those

discussed in chapter 4, for Spanish and Basque, are being currently collected according to this methodology.

In the future, we should consider whether there are ways to optimize data collection in on-line experiments. Studies of this kind differ from experiments conducted in laboratory settings because, in the first case, data can be collected for long periods of time and there does not have to be a well-defined ending to the data-collection process. On the other hand, when people volunteer to participate, it is also very difficult to predict how many participants will take part. Therefore, presenting items randomly may be suboptimal for some Web-based experiments and we may need to change the number of stimuli for which we can realistically collect data depending on the rate at which data is being collected. Optimization is particularly interesting if only one aspect of the data is of primary focus. For instance, if the goal is to obtain accurate average response times, it may be useful to adjust the number of presentations of a stimulus depending on whether it has reached statistical goals set forward by research objectives. The number of presentations of stimuli may also depend on the demographic characteristics of a subject. For instance, for members of groups what are difficult to recruit, items which are most probable to have different response characteristics in that group compared to other groups may be presented more often. However, it remains to be investigated to what extent such optimization would affect usability of the data for other purposes. For example, the quality of the accuracy data could suffer because of the optimization for the collection of the reaction times.

COMPUTATION, ANALYSIS, AND MODELING USING LARGE DATASETS

It has been suggested that the increasing usage of large datasets in psychology will encourage an application of more sophisticated analytical and computational techniques and that psychoinformatics will emerge as a new sub-discipline of psychology (Yarkoni, 2012). We indeed experienced that working with such datasets encourages or even enforces usage of advanced computational methods. This is true with respect to collection, cleaning, and analysis of text corpora, as it requires programming web crawlers and near duplicate detection tools as well as implementing distributional semantics methods. Making the data available for other researchers often means developing a web interface that allows access to the dataset. Collecting on-line megastudy data requires programming the web based experiments as well as operating servers, often under a heavy-load. Finally, the analysis of large datasets inspires creativity and the application of machine learning techniques.

Even more importantly, the availability of the large datasets can also facilitate the development of computational models. In chapter 5, I discussed how connectionist models of distributional semantics (Mikolov, Chen, Corrado, & Dean, 2013) share many properties with well-established models of learning. This is an interesting situation as connectionist models such as these blend state-of-the art computational techniques with implementational principles that could form a good basis for cognitive modeling. This is even more important as the availability of large amounts of data and the increase in available computational power has led to a renaissance of connectionist modeling under the alias of deep learning (see LeCun,

Bengio, & Hinton, 2015 for a review) that often provide cutting-edge performance in artificial intelligence tasks. Of course, basing a model on connectionist ideas does not automatically make it plausible as a cognitive model. However, many of the deep learning techniques are biologically inspired, making it worthwhile to consider them as a good basis for computational models in psychology.

An example of a field in psycholinguistics that could benefit from incorporating deep learning in combination with large text corpora is the modeling of orthographic representations. It is known that the visual cortex is organized into a hierarchy of feature detectors that are sensitive to progressively complex patterns in the visual input (Hubel & Wiesel, 1962) and it has been suggested that, in addition to normal reading, such an organization explains our ability to read a text with scrambled letters with relative ease, and transposed letter priming (Dehaene, Cohen, Sigman, & Vinckier, 2005). Convolutional neural networks, which are structured in a way that closely resembles the organization of the visual cortex, have recently proved to excel at various engineering tasks and technical frameworks have been created that could allow to implement and train such networks on large text corpora. A comprehensive evaluation of such models should be relatively easy using datasets such as orthographic priming megastudies (Adelman et al., 2014). As a result, distributed representations of the orthographic code based on biologically plausible principles, could be developed in a similar way to what was discussed in chapter 5 in the context of distributional semantics.

REFERENCES

Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., … others. (2014). A behavioral database for masked form priming. *Behavior Research Methods*, *46*(4), 1052–1067.

Baayen, R. H., Milin, P., Filipovic Durdevic, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review 118*, 438-482.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459.

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods*, *44*(4), 998–1006. http://doi.org/10.3758/s13428-012-0195-z

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, *8*(3), e57410. http://doi.org/10.1371/journal.pone.0057410

Damian, M. F. (2010). Does variability in human performance outweigh imprecision in response devices such as computer keyboards? *Behavior Research Methods*, *42*(1), 205–211. http://doi.org/10.3758/BRM.42.1.205

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: a proposal. *Trends in Cognitive Sciences*, *9*(7), 335–341. http://doi.org/10.1016/j.tics.2005.05.004

Hubel, D. H., Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(1): 106–154.

Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *The Quarterly Journal of Experimental Psychology*, *61*(7), 1036–1066. http://doi.org/10.1080/17470210701438111

Katz, S. M. (1996). Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, *2*(01), 15–59.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2011). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English

words. *Behavior Research Methods*, *44*(1), 287–304. http://doi.org/10.3758/s13428-011-0118-4

Markowetz, A., Błaszkiewicz, K., Montag, C., Switala, C., & Schlaepfer, T. E. (2014). Psycho-Informatics: Big Data shaping modern psychometrics. *Medical Hypotheses*, *82*(4), 405–411. http://doi.org/10.1016/j.mehy.2013.11.030

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. Retrieved from http://arxiv.org/abs/1301.3781

Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, *47*(2), 309–327. http://doi.org/10.3758/s13428-014-0471-1

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H., & Prokasy, W. F. (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Nederlandse samenvatting

PSYCHOLINGUÏSTIEK OP GROTE SCHAAL: DE COMBINATIE VAN TEKSTCORPORA, MEGASTUDIES, EN GEDISTRIBUEERDE SEMANTIEK IN HET ONDERZOEK NAAR MENSELIJKE TAALVERWERKING

Het doel van dit proefschrift is om, op basis van de toegenomen beschikbaarheid van digitale materialen en de online toegang tot grote populaties van deelnemers, nieuwe psycholinguïstische middelen te ontwikkelen, de methodologie voor het creëren van dergelijke middelen te verbeteren en om deze nieuwe ontwikkelingen te gebruiken om psycholinguïstische theorie te bevorderen.

Hoofdstuk 2 stel een nieuwe lijst van woordfrequenties voor op basis van ondertitels van Brits Engelse tv-programma's. De nieuwe woordfrequenties verklaren meer variantie in de lexicale beslissingstijden van het British Lexicon Project dan bestaande woordfrequenties op basis van het British National Corpus en dan bestaande woordfrequenties die voornamelijk gebaseerd zijn op ondertitels uit Amerikaanse films. Naast frequenties van woordvormen bevatten de ontwikkelde gegevens ook *part-of-speech*-specifieke woordfrequenties, frequentie van lemmata, specifieke frequenties voor kinderprogramma's, en frequenties van woordbigrammen. Onderzoekers die in een Brits Engelse context werken, krijgen zo toegang tot een ruim scala aan frequenties. Het hoofdstuk beschrijft ook een nieuwe schaal voor woordfrequentie, de Zipf schaal, die

sommige misverstanden over het woordfrequentie-effect kan voorkomen.

In het tweede empirische hoofdstuk ontwikkel ik woordfrequenties op basis van filmondertitels in het Pools. In twee lexicale beslissingsexperimenten vergelijk ik de nieuwe normen met woordfrequenties die afgeleid zijn van een Pools corpus dat voornamelijk geschreven materiaal bevat. Daarnaast onderzoek ik of de evaluatie van frequentienormen efficiënter kan gemaakt worden door (1) een optimale keuze van de stimuli in het experiment en (2) door een webgebaseerde manier om experimentele gegevens te verzamelen. De resultaten geven aan dat de woordfrequenties uit beide corpora een verschillend potentieel hebben voor het verklaren van menselijke gedrag in verschillende frequentiebereiken en dat corpora op basis van schriftelijke materiaal frequenties van formele woorden ernstig overschatten. Een aantal van deze bevindingen hebben implicaties voor toekomstig onderzoek waarin frequentieschattingen worden vergeleken. Naast frequenties voor woordvormen bevatten de nieuwe normen ook de contextuele diversiteit van woordvormen, *part-of-speech*-specifieke woordfrequenties, frequenties van lemmata, en frequenties van woordbigrammen.

In hoofdstuk 4 richt ik me ook op het woordfrequentie-effect, maar maak ik gebruik van een online methode voor het verzamelen van gedragsgegevens op een veel grotere schaal. Hoewel woordfrequentielijsten meestal op logaritmische schaal gebruikt worden, is de relatie tussen log-getransformeerde frequenties en gedragsgegevens niet volledig lineair, maar vlakt ze af voor hoogfrequente woorden (e.g., Keuleers, Lacey, Rastle, & Brysbaert, 2012; Keuleers, Diependaele, & Brysbaert, 2010). Het is ook bekend

dat de grootte van het frequentie-effect afhankelijk is van de vaardigheid van de lezer (Diependaele, Lemhöfer & Brysbaert, 2012). In hoofdstuk 4 onderzoek ik of de voorgaande bevindingen kunnen verklaard worden door een combinatie van de statistische eigenschappen van woordfrequentiedistributies (gekenmerkt door een extreme verdeling en onderspecificatie van lage frequenties) met het door een power-functie beschreven leereffect (Newell & Rosenbloom, 1981). Op basis van simulaties met tekstcorpora en gedragsgegevens uit twee zeer grote experimenten in het Engels en in het Nederlands -meer dan anderhalf miljoen deelnemers met een ruime spreiding in demografische karakteristieken-, toon ik aan dat deze theoretische benadering meerdere verschijnselen in het onderzoek naar woordherkenning verklaart.

In de laatste twee empirische hoofdstukken, richt ik me op de vraag hoe informatie die uit tekstcorpora kan afgeleid worden door distributionele semantische technieken (e.g., Jones & Mewhort, 2007; Landauer & Dumais, 1997; Lund & Burgess, 1996) kan ingezet worden in psycholinguïstisch onderzoek. Recente ontwikkelingen op het gebied van distributionele semantiek (Mikolov et al., 2013) hebben geleid tot een nieuwe klasse van modellen die semantische gelijkenis tussen woorden uitdrukken aan de hand van een voorspellingsgebaseerde architectuur. In hoofdstuk 5 bespreek ik de relevantie van deze modellen voor psycholinguïstische theorieën en vergelijk ik ze met de meer traditionele distributionele semantische modellen zoals HAL. Daarna vergelijk ik de prestaties van de modellen op een grote dataset van semantische priming (Hutchison et al., 2013) en op een aantal andere semantische verwerkingstaken en besluit ik dat de voorspellingsgebaseerde modellen de gedragsdata

doorgaans beter verklaren. Op theoretisch vlak stel ik dat deze modellen de kloof dichten tussen de traditionele benaderingen van distributionele semantiek en psychologisch plausibele leerprincipes. Als hulpmiddel voor onderzoekers, stel ik voor een reeks van verschillende modellen semantische vectoren ter beschikking voor het Engels en voor het Nederlands en ontwikkel ik een gebruiksvriendelijke interface waarmee verschillende maten van semantische gelijkenis kunnen berekend worden.

In het laatste empirische hoofdstuk van dit proefschrift richt ik me op subjectieve waarderingen (ratings) voor variabelen zoals verwervingsleeftijd, concreetheid, en affectieve valentie. Subjectieve waarderingen zijn een belangrijk onderdeel van psycholinguïstisch onderzoek maar dekken, zelfs voor goed bestudeerde talen, vaak een klein deel van de woordenschat. Een mogelijke oplossing hiervoor omvat het gebruik van corpora om een semantische gelijkenisruimte te bouwen en vervolgens de toepassing van machine learning technieken om, op basis van bestaande data, ratings voor nieuwe woorden te extrapoleren. In hoofdstuk 6 voer ik een systematische vergelijking uit van twee extrapolatie technieken: *k-nearest neighbors* en *random forest* in combinatie met semantische ruimtes op basis van van latente *semantische analyse* (Landauer & Dumais, 1997), *topic models* (Blei, Ng, Jordanië, 2003), een versie van *hyperspace analoge of language* (HAL, Lund & Burgess, 1996), en een *skip-gram model* (Mikolov et al., 2013). Een variant van de *k-nearest-neighbors* methode met *skip-gram* vectoren doet de meest accurate voorspellingen, maar de *random forest* methode heeft het voordeel dat ze eenvoudig extra predictoren kan opnemen. Ik evalueer het nut van de methoden door na te gaan hoe goed menselijke prestaties in een lexicale

beslissingstaak kunnen worden verklaard door geëxtrapoleerde ratings voor verwervingsleeftijd en hoe precies woorden op basis van geëxtrapoleerde ratings kunnen toegewezen worden aan discrete categorieën. Ik merk op dat extrapolatiemethoden tot statistische artefacten kunnen leiden en dat in experimenten die gebruik maken van geëxtrapoleerde ratings andere conclusies bereikt kunnen worden dan in experimenten waar menselijke ratings worden gebruikt. Vanuit praktisch oogpunt is het nut van de met de beschreven methodes geëxtrapoleerde ratings daarom beperkt.

In het laatste hoofdstuk bespreek ik de praktische, methodologische en theoretische implicaties van dit proefschrift en doe ik een aantal suggesties voor toekomstig onderzoek.

REFERENTIES

Diependaele, K., Lemhöfer, K., & Brysbaert, M. (2013). The word frequency effect in first- and second-language word recognition: A lexical entrenchment account. *The Quarterly Journal of Experimental Psychology*, *66*(5), 843–863. http://doi.org/10.1080/17470218.2012.720994

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C.-S., … Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*(4), 1099–1114. http://doi.org/10.3758/s13428-012-0304-z

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37. http://doi.org/10.1037/0033-295X.114.1.1

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice Effects in Large-Scale Visual Word Recognition Studies: A Lexical Decision Study on 14,000 Dutch Mono- and Disyllabic Words and Nonwords. *Frontiers in Psychology*, *1*. http://doi.org/10.3389/fpsyg.2010.00174

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2011). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. http://doi.org/10.3758/s13428-011-0118-4

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211-240.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. Retrieved from http://arxiv.org/abs/1301.3781

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

Appendix: Data storage fact sheets

```
% Data Storage Fact Sheet

% Name/identifier study: BBC word frequency measures
% Author: Paweł Mandera
% Date: 2016-02-29


1. Contact details
============================================================

1a. Main researcher
------------------------------------------------------------

- name: Paweł Mandera
- address: Department of Experimental Psychology, Henri Dunantlaan 2, 9000
  Ghent, Belgium
- e-mail: pawel.mandera@ugent.be, pawel@pawelmandera.com

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------

- name: Marc Brysbaert
- address: Department of Experimental Psychology, Henri Dunantlaan 2, 9000
  Gent, Belgium
- e-mail: marc.brysbaert@ugent.be

If a response is not received when using the above contact details, please send
an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology
and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.


2. Information about the datasets to which this sheet applies
============================================================

* Reference of the publication in which the datasets are reported:

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014).
SUBTLEX-UK: A new and improved word frequency database for British
English. Quarterly Journal of Experimental Psychology, 67(6), 1176-1190.
(chapter 2)

* Which datasets in that publication does this sheet apply to?:

the corpus used to calculate frequency statistics, word frequencies


3. Information about the files that have been stored
============================================================


3a. Raw data
------------------------------------------------------------

* Have the raw data been stored by the main researcher? [ ] YES / [X] NO
```

If NO, please justify:

To respect the intellectual property rights of the British Broadcasting Corporation (BBC) the full textual content of the relevant subtitles was not stored or reproduced for the purpose of this research. For more information see Van Heuven, Mandera, Keuleers, & Brysbaert (2014), page 3.

* On which platform are the raw data stored?
  - [ ] researcher PC
  - [ ] research group file server
  - [ ] other (specify): ...

* Who has direct access to the raw data (i.e., without intervention of another person)?
  - [ ] main researcher
  - [ ] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

3b. Other files
---

* Which other files have been stored?
  - [ ] file(s) describing the transition from raw data to reported results. Specify: ...
  - [X] file(s) containing processed data. Specify: word frequencies derived from the subtitles are publicly available at http://crr.ugent.be/archives/1423
  - [ ] file(s) containing analyses. Specify: ...
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...

* On which platform are these other files stored?
  - [X] individual PC
  - [X] research group file server
  - [ ] other: ...

* Who has direct access to these other files (i.e., without intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [X] all members of the research group
  - [X] all members of UGent
  - [X] other (specify): dataset is publicly available

4. Reproduction
===

```
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
   - e-mail:


v0.2
```

% Data Storage Fact Sheet

% Name/identifier study: Raw corpus of Polish movie subtitles
% Author: Paweł Mandera
% Date: 2016-02-29

1. Contact details
==========

1a. Main researcher

- name: Paweł Mandera
- address: Department of Experimental Psychology, Henri Dunantlaan 2, 9000 Gent, Belgium
- e-mail: pawel.mandera@ugent.be, pawel@pawelmandera.com

1b. Responsible Staff Member (ZAP)

- name: Marc Brysbaert
- address: Department of Experimental Psychology, Henri Dunantlaan 2, 9000 Ghent, Belgium
- e-mail: marc.brysbaert@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
==========

* Reference of the publication in which the datasets are reported:

Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2015). SUBTLEX-PL: Subtitle-based word frequency estimates for Polish. Behavior Research Methods, 47(2), 471-483. (chapter 3)

* Which datasets in that publication does this sheet apply to?:

the corpus of movie subtitles used in the paper

3. Information about the files that have been stored
==========

3a. Raw data
----

* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
If NO, please justify:

```
* On which platform are the raw data stored?
  - [X] researcher PC
  - [X] research group file server
  - [ ] other (specify): ...

* Who has direct access to the raw data (i.e., without intervention of another
  person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


3b. Other files
------------------------------------------------------------

* Which other files have been stored?
  - [X] file(s) describing the transition from raw data to reported results.
    Specify: scripts used to clean the corpus
  - [X] file(s) containing processed data. Specify: word frequencies derived
    from this corpus are publicly available at
    http://crr.ugent.be/programs-data/subtitle-frequencies/subtlex-pl, tagged
    version of the corpus
  - [ ] file(s) containing analyses. Specify: ...
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how this
    content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...


* On which platform are these other files stored?
  - [X] individual PC
  - [ ] research group file server
  - [ ] other: ...

* Who has direct access to these other files (i.e., without intervention of
  another person)?
  - [X] main researcher
  - [ ] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


4. Reproduction
============================================================

* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
```

```
   - affiliation:
   - e-mail:


v0.2
```

```
% Data Storage Fact Sheet

% Name/identifier study: Subtlex-pl validation data
% Author: Paweł Mandera
% Date: 2016-02-29


1. Contact details
==========================================================

1a. Main researcher
----------------------------------------------------------

- name: Paweł Mandera
- address: Department of Experimental Psychology, Henri Dunantlaan 2, 9000
  Gent, Belgium
- e-mail: pawel.mandera@ugent.be, pawel@pawelmandera.com

1b. Responsible Staff Member (ZAP)
----------------------------------------------------------

- name: Marc Brysbaert
- address: Department of Experimental Psychology, Henri Dunantlaan 2, 9000
  Ghent, Belgium
- e-mail: marc.brysbaert@ugent.be

If a response is not received when using the above contact details, please send
an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology
and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.


2. Information about the datasets to which this sheet applies
==========================================================

* Reference of the publication in which the datasets are reported:

  Mandera, P., Keuleers, E., Wodniecka, Z., & Brysbaert, M. (2014). Subtlex-pl:
  subtitle-based word frequency estimates for Polish. Behavior Research
  Methods, 67(6), 1176-1190. http://doi.org/10.3758/s13428-014-0489-4
  (chapter 3)

* Which datasets in that publication does this sheet apply to?:

  Behavioral data used to validate the frequency norms in experiment 1 and 2.

3. Information about the files that have been stored
==========================================================


3a. Raw data
----------------------------------------------------------

* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
If NO, please justify:
```

```
* On which platform are the raw data stored?
  - [X] researcher PC
  - [X] research group file server
  - [ ] other (specify): ...

* Who has direct access to the raw data (i.e., without intervention of another
  person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


3b. Other files
------------------------------------------------------------

* Which other files have been stored?
  - [X] file(s) describing the transition from raw data to reported results.
    Specify: cleaning of the behavioral data, regression analyses in study 1 and
2
  - [ ] file(s) containing processed data. Specify: ...
  - [ ] file(s) containing analyses. Specify: ...
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how this
    content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...


* On which platform are these other files stored?
  - [X] individual PC
  - [ ] research group file server
  - [ ] other: ...

* Who has direct access to these other files (i.e., without intervention of
  another person)?
  - [X] main researcher
  - [ ] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


4. Reproduction
============================================================

* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
   - e-mail:
```

v0.2

```
% Data Storage Fact Sheet

% Name/identifier study: Raw corpus of English movie subtitles
% Author: Paweł Mandera
% Date: 2016-02-29


1. Contact details
============================================================

1a. Main researcher
------------------------------------------------------------

- name: Paweł Mandera
- address: Department of Experimental Psychology, Henri Dunantlaan 2, 9000
  Ghent, Belgium
- e-mail: pawel.mandera@ugent.be, pawel@pawelmandera.com

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------

- name: Marc Brysbaert
- address: Department of Experimental Psychology, Henri Dunantlaan 2, 9000
  Gent, Belgium
- e-mail: marc.brysbaert@ugent.be

If a response is not received when using the above contact details, please send
an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology
and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.


2. Information about the datasets to which this sheet applies
============================================================

* Reference of the publication in which the datasets are reported:

Mandera, P., Keuleers, E., & Brysbaert, M. An exposure-based account of the
changes in the word frequency effect. (chapter 4)

Mandera, P., Keuleers, E., & Brysbaert, M. (in press). Explaining human
performance in psycholinguistic tasks with models of semantic similarity based
on prediction and counting: A review and empirical validation.  Journal of
Memory and Language. (chapter 5)

Mandera, P., Keuleers, E., & Brysbaert, M. (2015). How useful are corpus-based
methods for extrapolating psycholinguistic variables?  Quarterly Journal of
Experimental Psychology, 68(8), 1623-1642. (chapter 6)

* Which datasets in that publication does this sheet apply to?:

the text corpora used to calculate word frequencies for English in chapter 4 and
to train distributional semantics models in chapter 5 and 6

3. Information about the files that have been stored
============================================================
```

3a. Raw data
----

* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
If NO, please justify:

* On which platform are the raw data stored?
  - [X] researcher PC
  - [X] research group file server
  - [ ] other (specify): ...

* Who has direct access to the raw data (i.e., without intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

3b. Other files
----

* Which other files have been stored?
  - [X] file(s) describing the transition from raw data to reported results. Specify: scripts used to clean the corpus, to calculate word frequencies and to train distributional semantics models in chapter 5
  - [ ] file(s) containing processed data. Specify: ...
  - [ ] file(s) containing analyses. Specify: ...
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...

* On which platform are these other files stored?
  - [X] individual PC
  - [ ] research group file server
  - [X] other: the semantic spaces trained using the text corpus are publicly available at http://crr.ugent.be/snaut/

* Who has direct access to these other files (i.e., without intervention of another person)?
  - [X] main researcher
  - [ ] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction
==

```
* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
   - e-mail:


v0.2
```

```
% Data Storage Fact Sheet

% Name/identifier study: Raw corpus of Dutch movie subtitles
% Author: Paweł Mandera
% Date: 2016-02-29


1. Contact details
============================================================

1a. Main researcher
------------------------------------------------------------

- name: Paweł Mandera
- address: Department of Experimental Psychology, Henri Dunantlaan 2, 9000
  Ghent, Belgium
- e-mail: pawel.mandera@ugent.be, pawel@pawelmandera.com

1b. Responsible Staff Member (ZAP)
------------------------------------------------------------

- name: Marc Brysbaert
- address: Department of Experimental Psychology, Henri Dunantlaan 2, 9000
  Gent, Belgium
- e-mail: marc.brysbaert@ugent.be

If a response is not received when using the above contact details, please send
an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology
and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.


2. Information about the datasets to which this sheet applies
============================================================

* Reference of the publication in which the datasets are reported:

Mandera, P., Keuleers, E., & Brysbaert, M. An exposure-based account of the
changes in the word frequency effect. (chapter 4)

Mandera, P., Keuleers, E., & Brysbaert, M. (in press). Explaining human
performance in psycholinguistic tasks with models of semantic similarity based
on prediction and counting: A review and empirical validation.  Journal of
Memory and Language. (chapter 5)

* Which datasets in that publication does this sheet apply to?:

the text corpora used to calculate word frequencies for Dutch in chapter 4 and
to train distributional semantics models in chapter 5

3. Information about the files that have been stored
============================================================


3a. Raw data
------------------------------------------------------------
```

* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
If NO, please justify:

* On which platform are the raw data stored?
  - [X] researcher PC
  - [X] research group file server
  - [ ] other (specify): ...

* Who has direct access to the raw data (i.e., without intervention of another person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

3b. Other files
---

* Which other files have been stored?
  - [X] file(s) describing the transition from raw data to reported results. Specify: scripts used to clean the corpus, to calculate word frequencies and to train distributional semantics models
  - [X] file(s) containing processed data. Specify: derived word frequencies, distributional semantics models
  - [ ] file(s) containing analyses. Specify: ...
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how this content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...

* On which platform are these other files stored?
  - [X] individual PC
  - [ ] research group file server
  - [X] other: the semantic spaces trained using the text corpus are publicly available at http://crr.ugent.be/snaut/

* Who has direct access to these other files (i.e., without intervention of another person)?
  - [X] main researcher
  - [ ] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...

4. Reproduction
===

* Have the results been reproduced independently?: [ ] YES / [X] NO

```
* If yes, by whom (add if multiple):
   - name:
   - address:
   - affiliation:
   - e-mail:


v0.2
```

% Data Storage Fact Sheet

% Name/identifier study: An exposure-based account of the changes in the word frequency effect -- vocabulary tests data
% Author: Paweł Mandera
% Date: 2016-02-29

1. Contact details
==========

1a. Main researcher

- name: Paweł Mandera
- address: Department of Experimental Psychology, Henri Dunantlaan 2, 9000 Gent, Belgium
- e-mail: pawel.mandera@ugent.be, pawel@pawelmandera.com

1b. Responsible Staff Member (ZAP)

- name: Marc Brysbaert
- address: Department of Experimental Psychology, Henri Dunantlaan 2, 9000 Gent, Belgium
- e-mail: marc.brysbaert@ugent.be

If a response is not received when using the above contact details, please send an email to data.pp@ugent.be or contact Data Management, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, 9000 Ghent, Belgium.

2. Information about the datasets to which this sheet applies
==========

* Reference of the publication in which the datasets are reported:

Mandera, P., Keuleers, E., & Brysbaert, M. An exposure-based account of the changes in the word frequency effect. (chapter 3)

* Which datasets in that publication does this sheet apply to?:

All vocabulary test data used in the analyses in chapter 4.

3. Information about the files that have been stored
==========

3a. Raw data
----

* Have the raw data been stored by the main researcher? [X] YES / [ ] NO
If NO, please justify:

```
* On which platform are the raw data stored?
  - [X] researcher PC
  - [X] research group file server
  - [ ] other (specify): ...

* Who has direct access to the raw data (i.e., without intervention of another
  person)?
  - [X] main researcher
  - [X] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


3b. Other files
------------------------------------------------------------

* Which other files have been stored?
  - [X] file(s) describing the transition from raw data to reported results.
    Specify: sql files merging of database tables (optimized for data
    collection) into processed data files (optimized for analyses), cleanining
    of the datasets
  - [X] file(s) containing processed data. Specify: merged and cleaned
    datasets, results of sampling and aggregation
  - [X] file(s) containing analyses. Specify: sampling and model fitting files
  - [ ] files(s) containing information about informed consent
  - [ ] a file specifying legal and ethical provisions
  - [ ] file(s) that describe the content of the stored files and how this
    content should be interpreted. Specify: ...
  - [ ] other files. Specify: ...


* On which platform are these other files stored?
  - [X] individual PC
  - [ ] research group file server
  - [ ] other: ...

* Who has direct access to these other files (i.e., without intervention of
  another person)?
  - [X] main researcher
  - [ ] responsible ZAP
  - [ ] all members of the research group
  - [ ] all members of UGent
  - [ ] other (specify): ...


4. Reproduction
============================================================

* Have the results been reproduced independently?: [ ] YES / [X] NO

* If yes, by whom (add if multiple):
   - name:
```

```
   - address:
   - affiliation:
   - e-mail:


v0.2
```